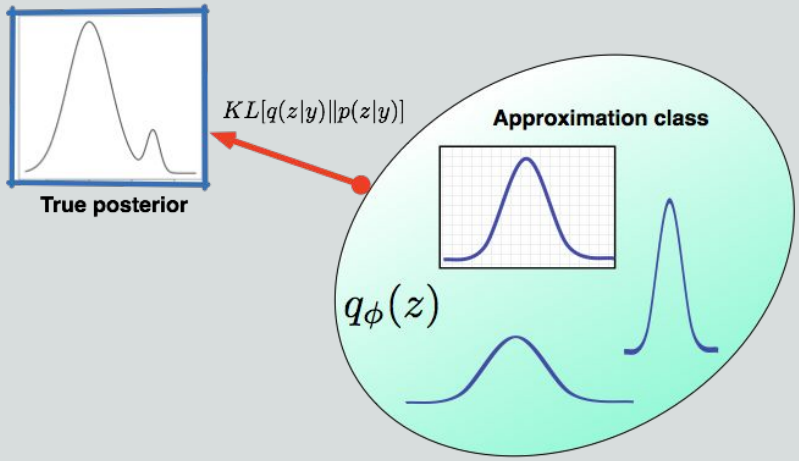
Variational Approximation



$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)]$$

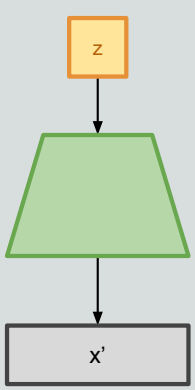
$$\log \mathbb{E}_{p(z)}[p_\theta(x_i|z)] = \log \mathbb{E}_{q_i(z)}\left[\frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right], \quad \forall q_i > 0$$

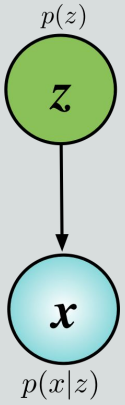
$$\log \mathbb{E}_{q_i(z)}\left[\frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right] \geqslant \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

$$\log p_\theta(\mathcal{D}) \geqslant \sum_{i=1}^{N} \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

Known as proposal, encoder or posterior model

Variational Inference: ELBO





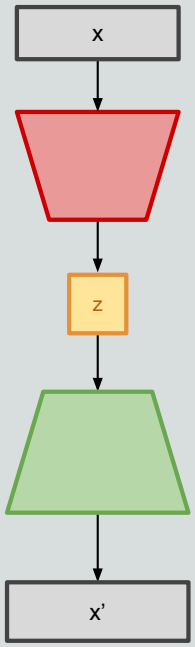
$$\log p_\theta(\mathcal{D}) \geqslant \sum_{i=1}^{N} \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

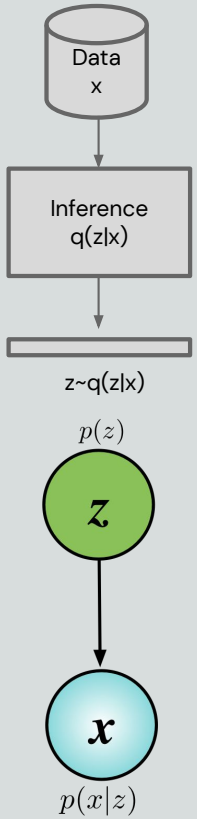
$$\mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right] = \mathbb{E}_{q_i(z)}[\log p_\theta(x_i|z)] - \mathrm{KLD}(q_i\|p)$$

Reconstruction (distortion)

Regularizer (rate)

Amortised Inference





Want to learn more?

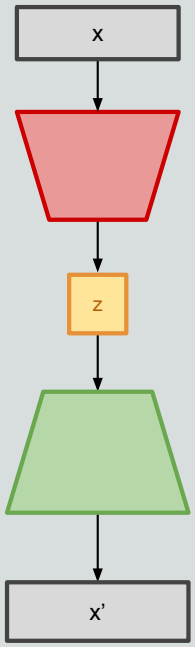
Stochastic Backpropagation and Approximate Inference in Deep Generative Models, Rezende et al, ICML 2014

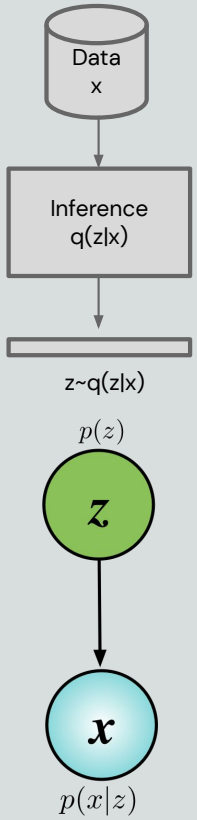
$$q_i^\star(z) = \text{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}[-\mathcal{F}(x_i, z)]$$

Introduce a parametric family of conditional densities

$$\text{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}[-\mathcal{F}(x_i, z)] \Rightarrow \text{argmax}_{\phi} \mathbb{E}_{q_\phi(z|x)}[-\mathcal{F}_\phi(x_i, z)$$

Variational AutoEncoder (VAE)





Want to learn more?

Stochastic Backpropagation and Approximate Inference in Deep Generative Models, Rezende et al, ICML 2014

Auto-Encoding Variational Bayes, Kingma & Welling, ICLR 2014

Deep Latent Gaussian Model p(x,z)

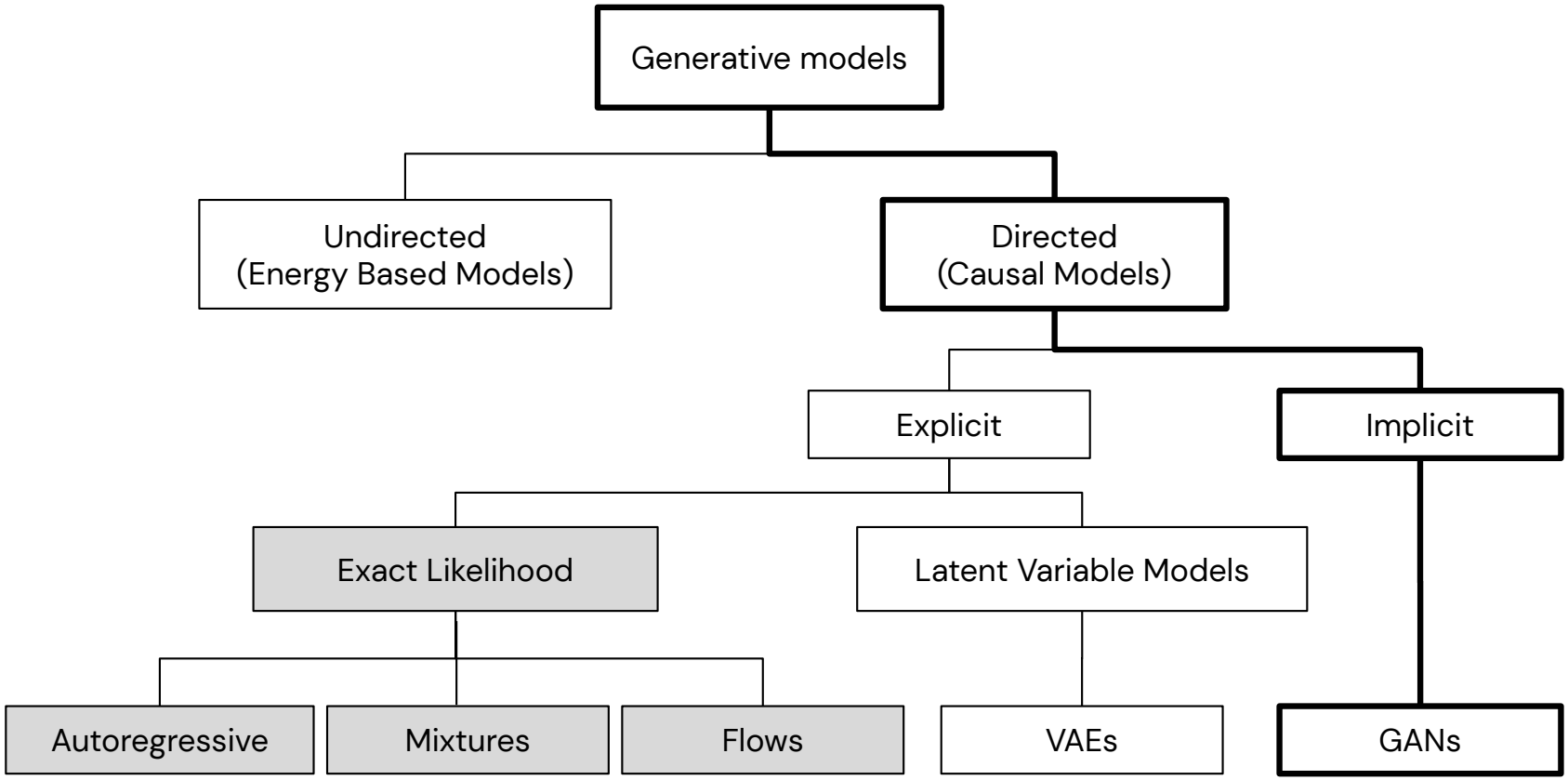
$$\begin{aligned} \text{prior sample} \quad & z \sim \mathcal{N}(0, \mathbb{I}) \\ \text{data sufficient statistics} \quad & \eta = f_\theta(z) \\ \text{data conditional likelihood} \quad & x \sim \mathcal{N}(\eta) \end{aligned}$$

Gaussian Recognition Model q(z)

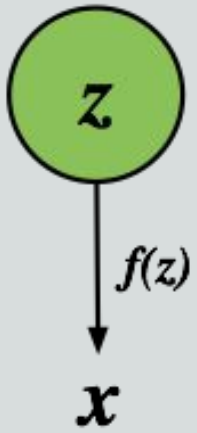
$$\begin{aligned} \text{data sample} \quad & x \sim \mathcal{D} \\ \text{latent sufficient statistics} \quad & \eta = f_\phi(x) \\ \text{posterior sample} \quad & z \sim \mathcal{N}(\eta) \end{aligned}$$

$$\mathbb{E}_{q_i(z)}[\log p_\theta(x_i|z)] - \mathrm{KLD}(q_i \| p)$$

Mapping out the landscape of Generative Models



Latent Variable Models



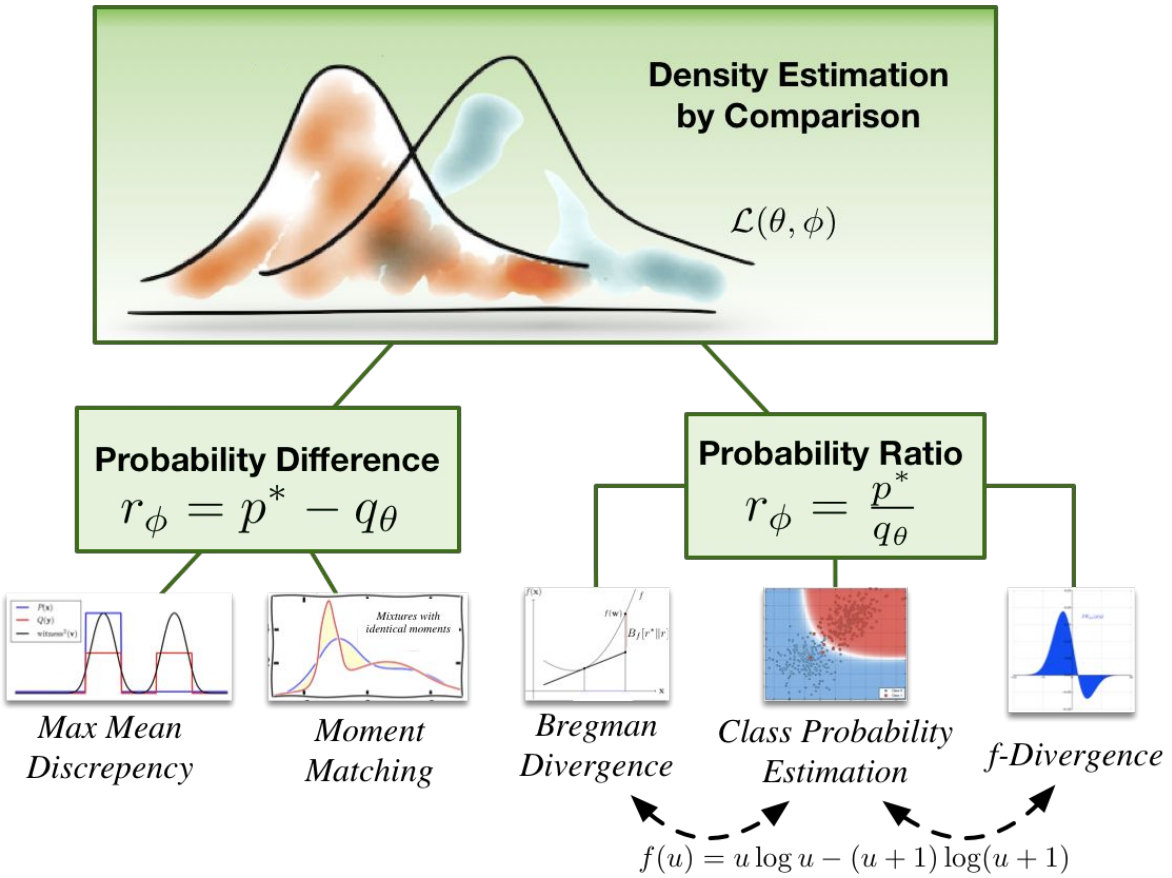
$$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$$
$$\mathcal{D} = \{x_i\} \quad i \in \{1, ..., N\}$$

$$\ln p_\theta(x) = \ln \int \delta(x - f(z)) p(z) dz$$

Learning by comparison

Want to learn more?

Learning in Implicit Generative Models, Mohamed and Lakshminarayanan, ICML 2017



Unsupervised-as-Supervised Learning

Want to learn more?

Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

Scoring Function

$$p(y=+1|\mathbf{x}) = D_\theta(\mathbf{x}) \qquad p(y=-1|\mathbf{x}) = 1 - D_\theta(\mathbf{x})$$

Bernoulli Loss

$$\mathcal{F}(\mathbf{x}, \theta, \phi) = \mathbb{E}_{p^*(x)}[\log D_\theta(\mathbf{x})] + \mathbb{E}_{q_\phi(x)}[\log(1 - D_\theta(\mathbf{x})]$$

Alternating optimisation

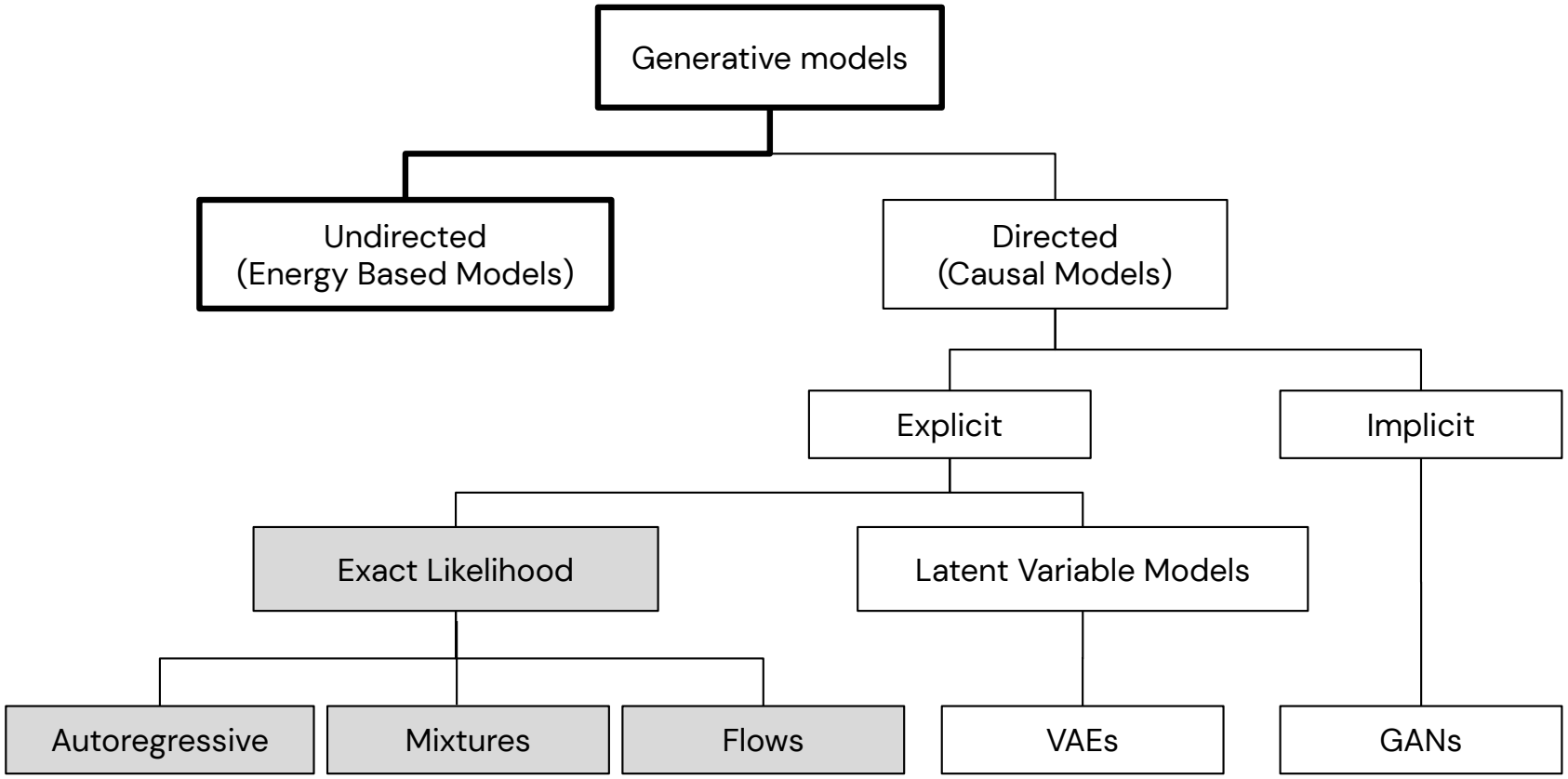
$$\min_\phi \max_\theta \mathcal{F}(\mathbf{x}, \theta, \phi)$$

- Use when we have differentiable simulators and models
- Can form the loss using any proper scoring rule.

Other names and places:

- Unsupervised and supervised learning
- Continuously updating inference
- Classifier ABC
- Generative Adversarial Networks

Mapping out the landscape of Generative Models

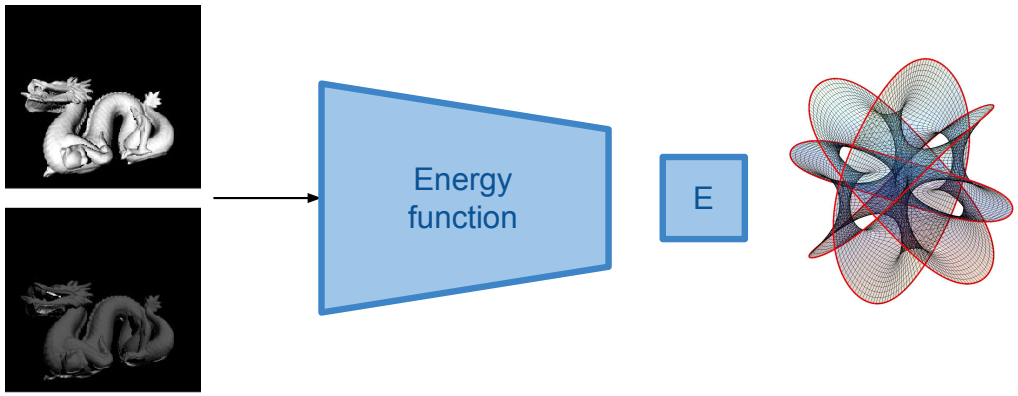


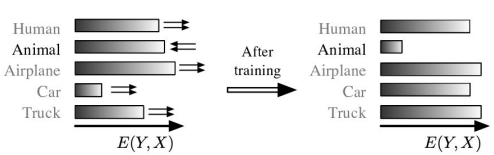
Energy models

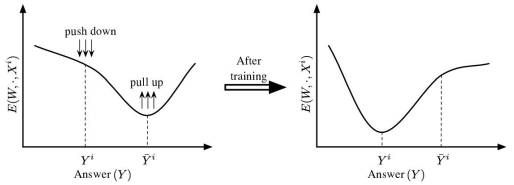
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples





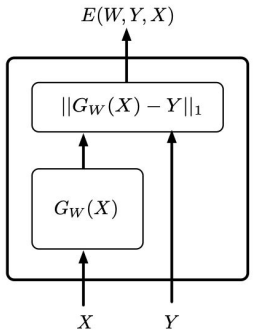


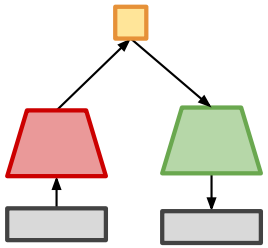
Energy models: $y = x^2$

Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

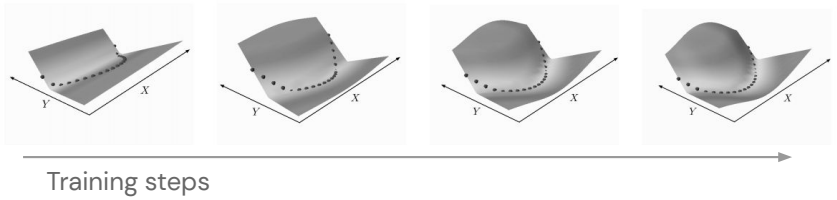
- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples





Energy: $E(W, Y, X) = ||G_W(X) - Y||_1.$

Loss: $\mathcal{L}_{\text{energy}}(W, \mathcal{S}) = \frac{1}{P}\sum_{i=1}^{P} E(W, Y^i, X^i) = \frac{1}{P}\sum_{i=1}^{P} ||G_W(X) - Y||_1.$

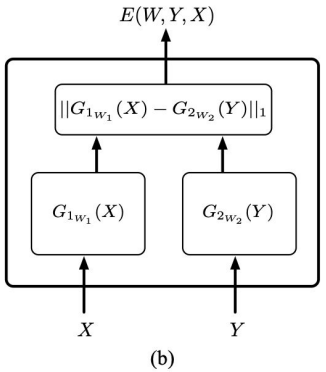


Energy models: $y = x^2$

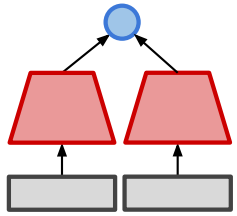
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples

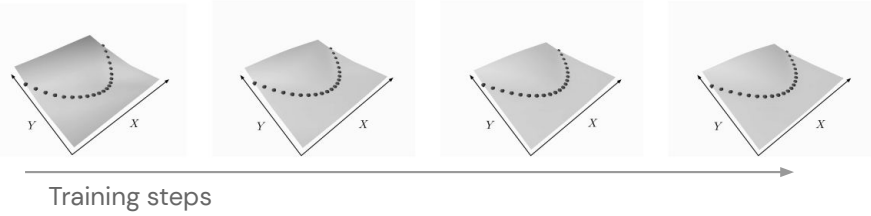


(b)



Energy: $E(W, X, Y) = ||G_{1_{W_1}}(X) - G_{2_{W_2}}(Y)||_1,$

Loss: $\mathcal{L}_{\text{energy}}(W, \mathcal{S}) = \frac{1}{P}\sum_{i=1}^{P} E(W, Y^i, X^i)$

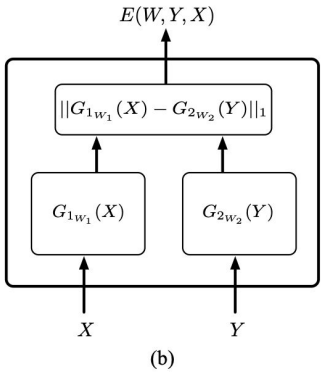


Energy models: $y = x^2$

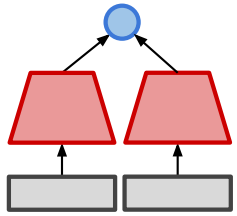
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples

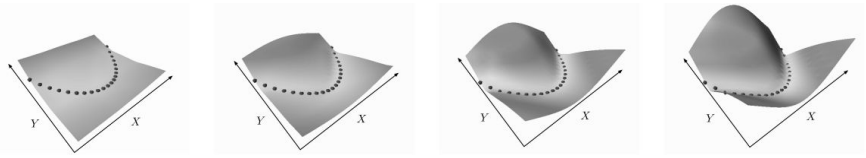


(b)



Energy: $E(W, X, Y) = ||G_{1_{W_1}}(X) - G_{2_{W_2}}(Y)||_1,$

Loss: $L(W, Y^i, X^i) = E(W, Y^i, X^i)^2 - \left(max(0, m - E(W, \bar{Y}^i, X^i))\right)^2.$

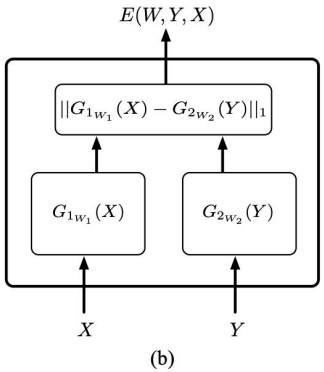


Energy models: $y = x^2$

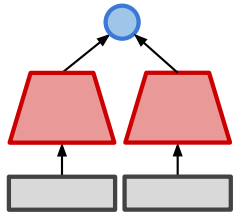
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples

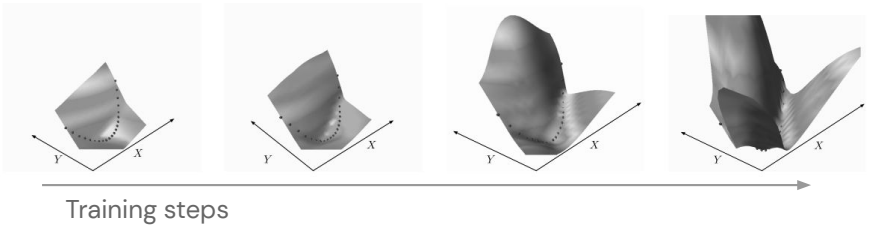


(b)



Energy: $E(W, X, Y) = ||G_{1_{W_1}}(X) - G_{2_{W_2}}(Y)||_1,$

Loss: $L(W, Y^i, X^i) = E(W, Y^i, X^i) + \frac{1}{\beta} \log \left(\int_{y \in \mathcal{Y}} e^{-\beta E(W, y, X^i)} \right).$



Maximum-Likelihood: Minimizing KL(data; model)

$$p(x;\theta) = Z(\theta)^{-1} e^{-E(x;\theta)}$$

$$Z(\theta) = \int dx e^{-E(x;\theta)}$$

$$\nabla_\theta \ln p(x;\theta) = -\nabla_\theta E(x;\theta) + \mathbb{E}_p[\nabla_\theta E(x';\theta)]$$

Easy:
evaluated at data

Hard:
evaluated at model samples

Sampling from energies: Langevin sampler

SDE: continuous time stochastic process

$$dX = -\nabla_x E(x;\theta)dt + \sqrt{2}d\xi$$

ξ is a stationary Gaussian process

$$\mathbb{E}[\xi(t)] = 0 \quad \mathbb{E}[\xi(t)\xi(t')] = \delta(t-t')$$

Basically, gradient descent with noise

$$x_{t+1} = x_t - \nabla_{x_t} E(x_t;\theta)dt + \sqrt{2dt}\xi$$

Will converge if **dt** → **0** and **t** → **∞**; but can easily get stuck in local minima

$$x_\infty \sim e^{-E(x)}$$

Langevin sampler: fixing discretisation errors

Want to learn more?

Optimal scaling of discrete approximations to Langevin diffusions, Roberts & Rosenthal, Journal of the Royal Statistical Society 2002

$$x_{t+1} = x_t - \nabla_{x_t} E(x_t;\theta)dt + \sqrt{2dt}\xi$$

$$x_{t+1} \sim q(x_{t+1}|x_t) = \mathbb{N}(x_{t+1}; x_t - \nabla_{x_t} E(x_t;\theta)dt, \sqrt{2dt})$$

Accept samples with probability

$$\alpha := \min\left\{1, \frac{q(x_t|x_{t+1})}{q(x_{t+1}|x_t)} e^{-E(x_{t+1})+E(x_t)}\right\}$$

Optimal dt must be chosen such that **E[α]=0.574**

Sampling from energies:
Hamiltonian Monte-Carlo sampler

$$H(x,p) = E(x) + K(p)$$

$$\frac{dx}{dt} = \frac{\partial H(x,p)}{\partial p} = \frac{\partial K(p)}{\partial p}$$

$$\frac{dp}{dt} = -\frac{\partial H(x,p)}{\partial x} = -\frac{\partial E(x)}{\partial x}$$

$$p(t=0) \sim \mathbb{N}(0, \mathbb{I})$$

$$(x_\infty, p_\infty) \sim e^{-H(x,p)} = e^{-E(x)-K(p)}$$

$$\Rightarrow x_\infty \sim e^{-E(x)}$$

Sampling from energies: Hamiltonian Monte-Carlo sampler

Want to learn more?

MCMC using Hamiltonian Dynamics, Neal, Handbook of Markov Chain Monte Carlo 2011

$$p(t + \frac{dt}{2}) = p(t)\frac{dt}{2}\nabla_x E(x)$$

Use leap-frog

$$x(t + dt) = x(t) + dt\nabla_{p(t+\frac{dt}{2})} K(p(t + \frac{dt}{2}))$$

$$p(t + dt) = p(t + \frac{dt}{2}) - \frac{dt}{2}\nabla_{x(t+dt)} E(x(t + dt))$$

Accept with probability $\alpha = \min\left\{1, e^{H(x,p)-H(x',p')}\right\}$

Score Matching:
Minimizing the Fisher divergence

$$q(x;\theta) = Z^{-1} e^{-E(x;\theta)}$$

$$Z(\theta) = \int dx e^{-E(x;\theta)}$$

Usually unknown

$$\begin{aligned}\mathrm{FD}(p,q) &= \mathbb{E}_p[\|\nabla_x \ln p(x) - \nabla_x \ln q(x;\theta)\|^2] \\ &= 2\mathbb{E}_p[\mathrm{Tr}[\nabla_x^2 \ln q(x;\theta)] + \|\nabla_x \ln q(x;\theta)\|^2] + \mathrm{cst}\end{aligned}$$

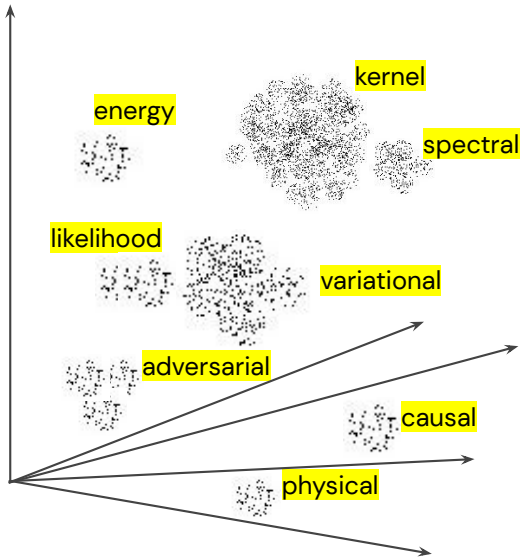
Integration by parts, **p**, **q** and their gradients must go to zero at the boundary

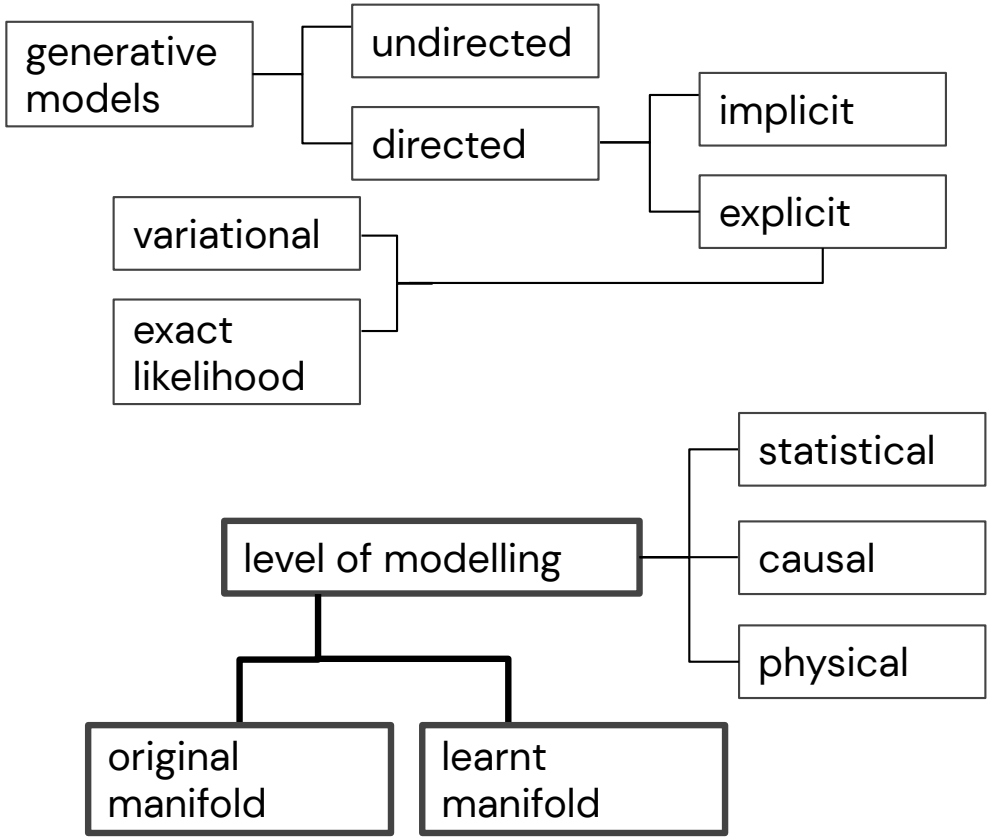
Easy: average over data

Expensive: Hessian

Good: Does not depend on normalizer **z**

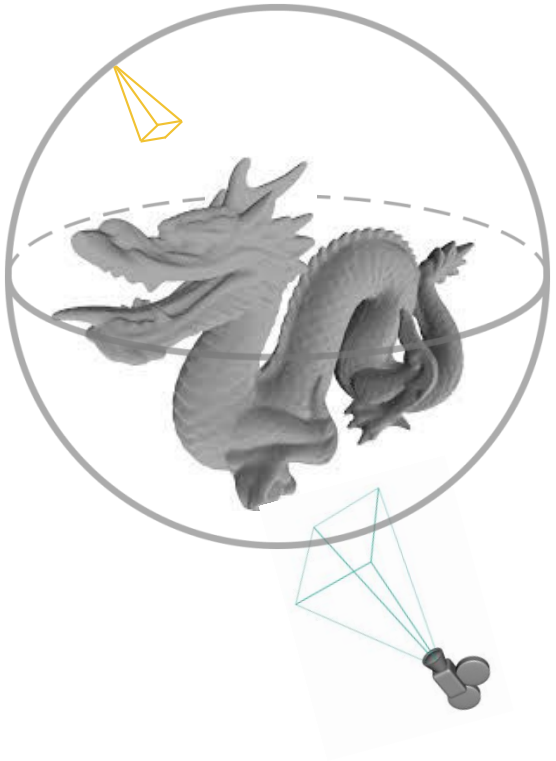
Mapping out the landscape



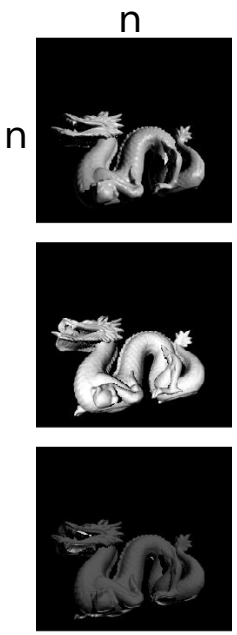


Manifold hypothesis

"Real-world high dimensional data lie on low-dimensional manifolds embedded in the high-dimensional space."

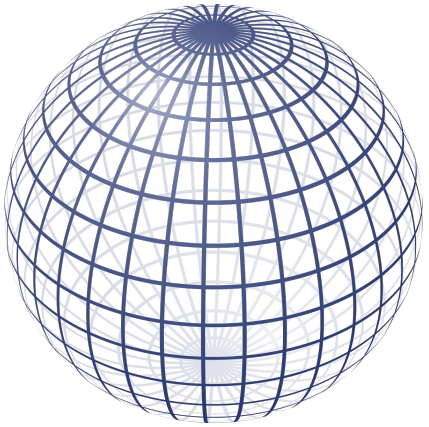


Pixel space



$\mathbb{R}^{n^2}$

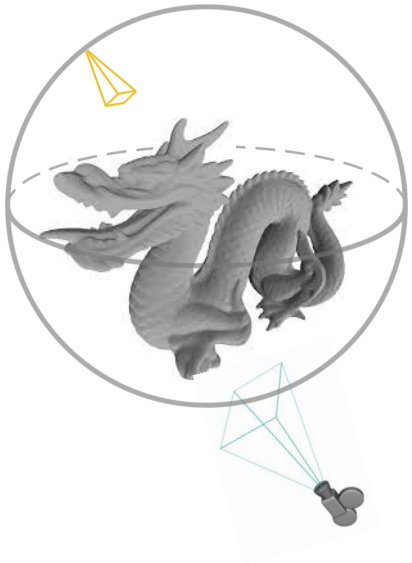
Sphere



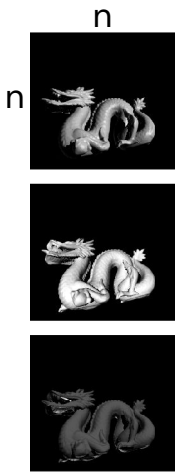
$\mathrm{S}^2 \in \mathbb{R}^3$

Manifold hypothesis

"Real-world high dimensional data lie on low-dimensional manifolds embedded in the high-dimensional space."

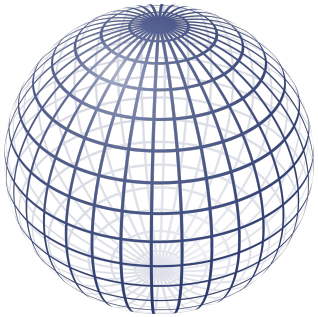


Pixel space

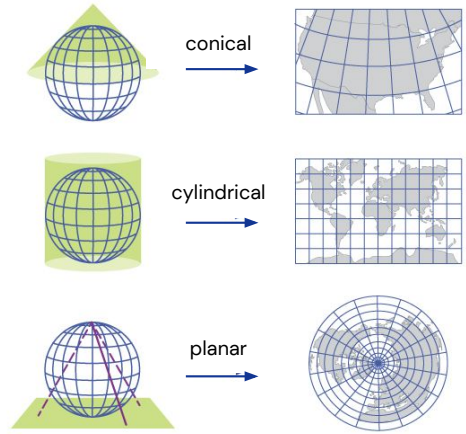


$\mathbb{R}^{n^2}$

Sphere



$\mathrm{S}^2 \in \mathbb{R}^3$



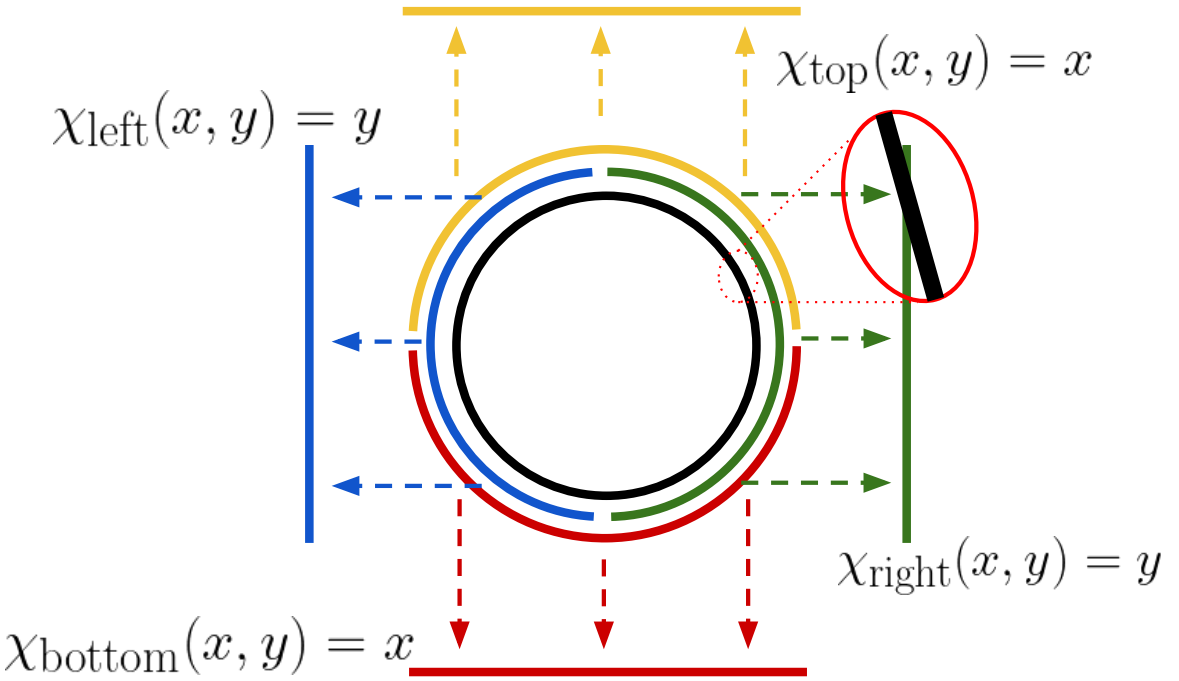
$\mathbb{R}^2$

Manifolds

Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Functions and open regions they map are called ***charts***
- Set of all charts that map the whole manifold make an ***atlas***



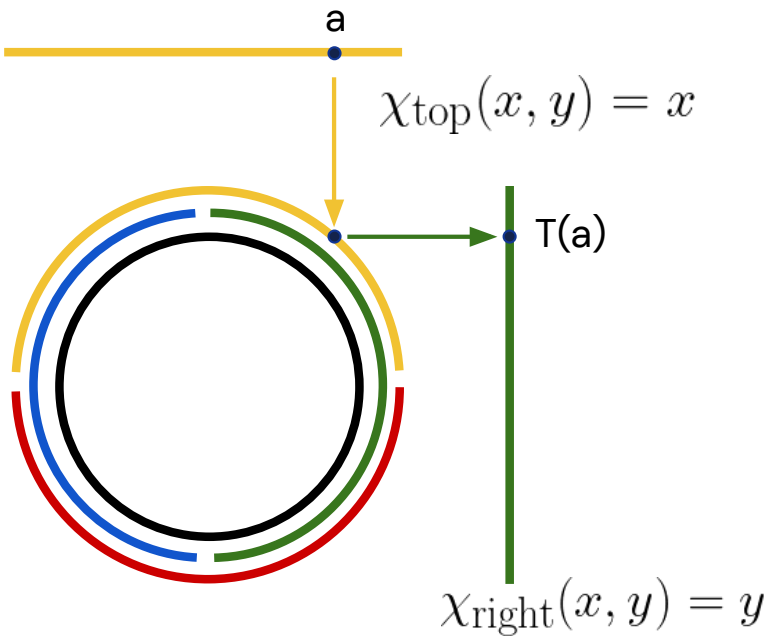
Manifolds

Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Functions and open regions they map are called ***charts***
- Set of all charts that map the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***

$$T : (0,1) \to (0,1) = \chi_{\text{right}} \circ \chi_{\text{top}}^{-1}$$



Manifolds

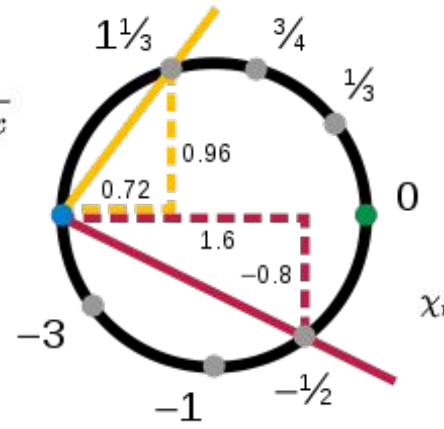
Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Continuous and invertible functions from segment to open region are called ***charts***
- Set of all charts that cover the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***
- Many possible choices for charts and atlases

$$\chi_{\text{plus}}(x, y) = t = \frac{y}{1 - x}$$

$$t = \frac{1}{s}$$



$$\chi_{\text{minus}}(x, y) = s = \frac{y}{1 + x}$$

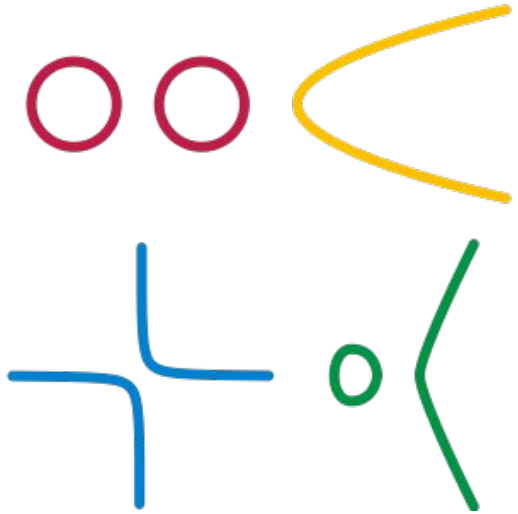
Manifolds

Want to learn more?

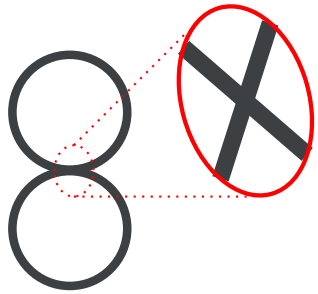
Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Homeomorphic maps from subset of manifold to Euclidean space R^n are called ***charts***
- Set of all charts that cover the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***
- Many possible choices for charts and atlases
- Manifolds need not be ***connected*** or ***closed***

1D Manifolds



Not a manifold



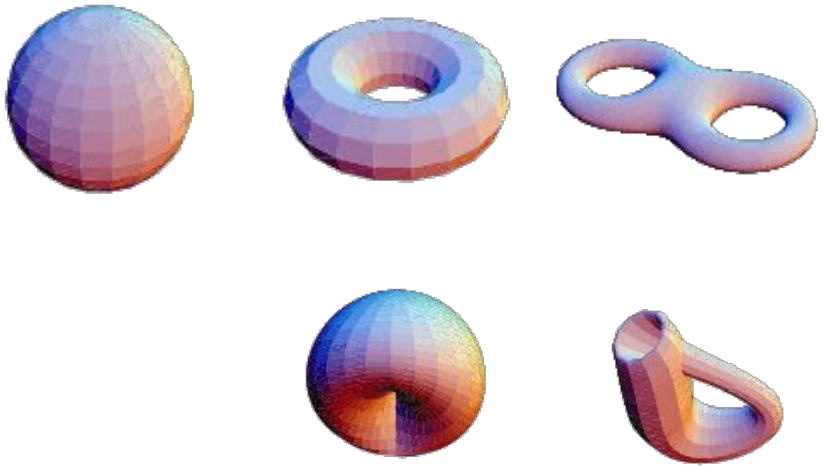
Manifolds

Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

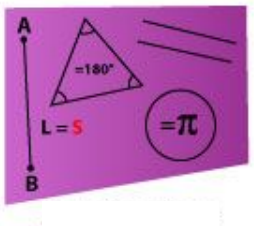
- Topological space that "locally" resembles Euclidean space.
- Homeomorphic maps from subset of manifold to Euclidean space R^n are called ***charts***
- Set of all charts that cover the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***
- Many possible choices for charts and atlases
- Manifolds need not be ***connected*** or ***closed***

2D Manifolds

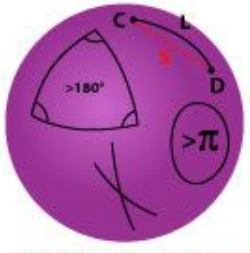


Euclidean space as a manifold

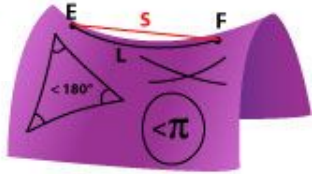
- R^n is a manifold
- Single chart = identity function
- Atlas contains single chart
- Distance = inner product



No curvature



Positive curvature



Negative curvature

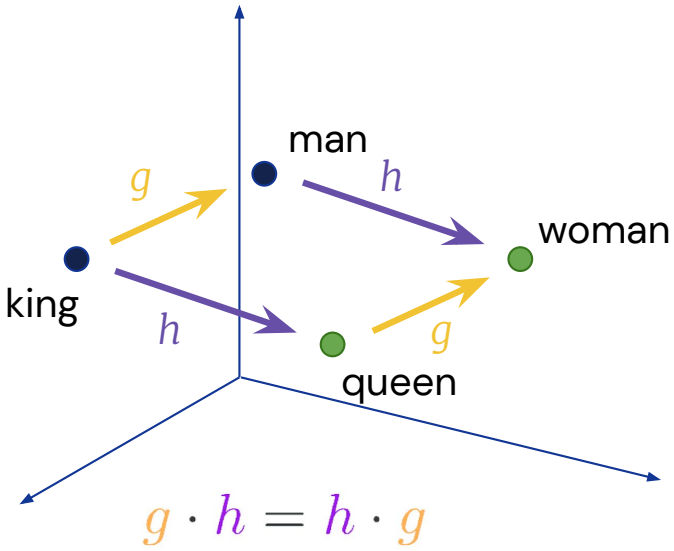
Euclidean space as a manifold

Want to learn more?

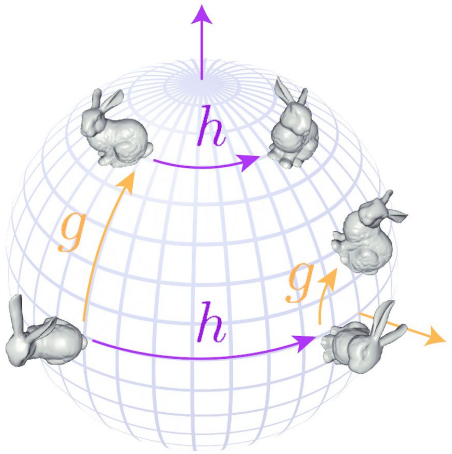
Efficient Estimation of Word Representations in Vector Space, Mikolov et al, ICLR 2013

Disentangling by Subspace Diffusion, Pfau et al, arxiv 2020

- R^n is a manifold
- Single chart = identity function
- Atlas contains single chart
- Distance = inner product



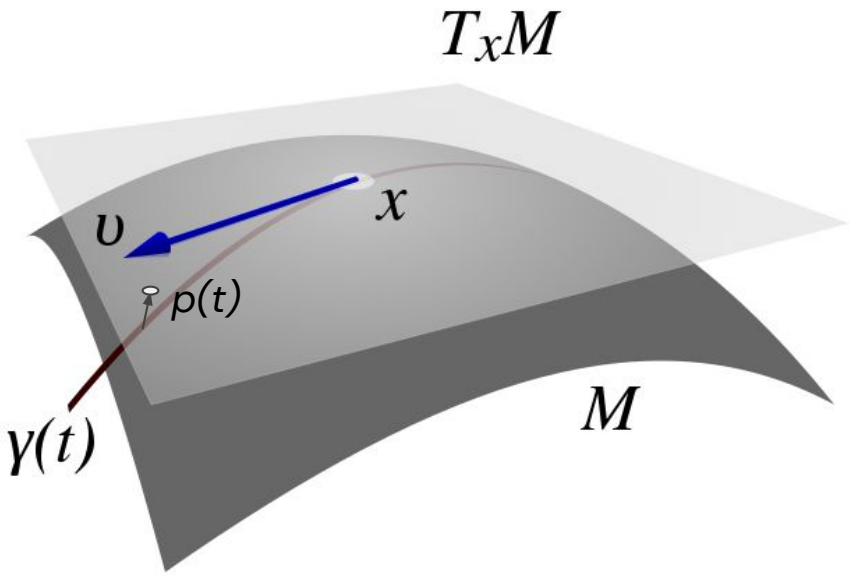
$$g \cdot h = h \cdot g$$



$$g \cdot h \neq h \cdot g$$

Tangent space

- Chart φ: **U** → $\mathbf{R}^n$ where **U** is an open subset of **M** containing **x**
- Curve γ: **t** → **M** runs along manifold through point **x**
- $\mathbf{p} = \varphi \circ \gamma(\mathbf{t})$ is position vector in $\mathbf{R}^n$
- $\mathbf{v} = d\mathbf{p}/d\mathbf{t} = (d\varphi \circ \gamma(\mathbf{t}))/d\mathbf{t}$, where $\mathbf{t} = \mathbf{t}_0$ is the "velocity" at **x**
- $\mathbf{v} \in \mathbf{R}^n$ is the tangent vector at **x**
- Tangent space $\mathbf{T}_x\mathbf{M}$ on manifold **M** at point **x** is the collection of all tangent vectors $\mathbf{v} \in \mathbf{R}^n$



Riemannian metric

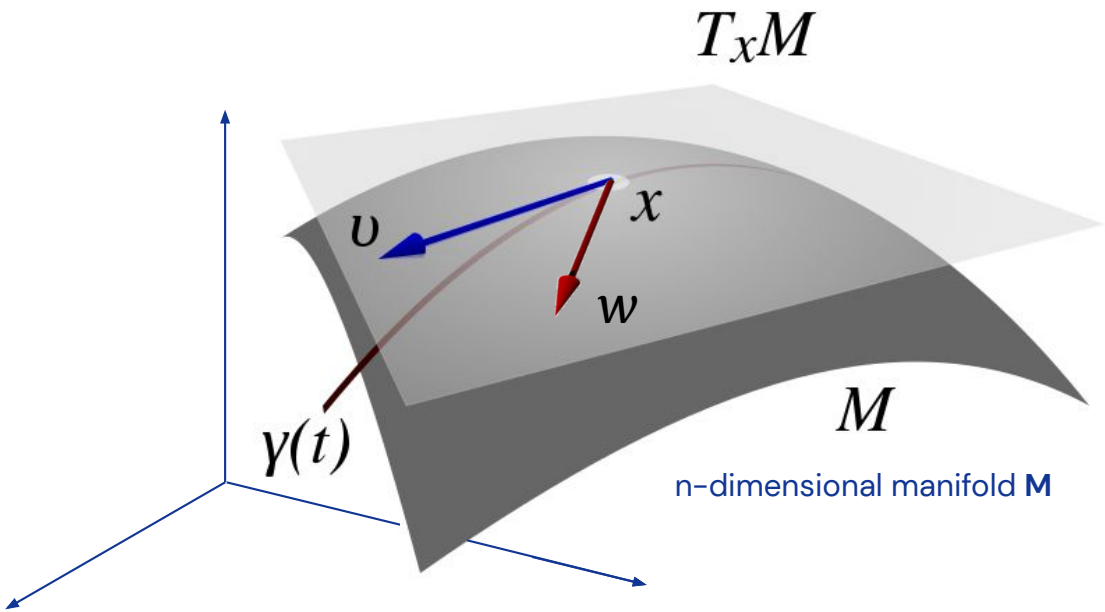
- Necessary to calculate distances (geodesics), angles, area etc.
- Riemannian metric is a family of function:

  $g_x : T_xM \times T_xM \rightarrow R, x \in M$

  such that $x \rightarrow g_x(v, w)$ is a smooth function of **x.**
- Generalisation of inner product:

  $g_x(v, w) = v^T J_q^T J_q w$

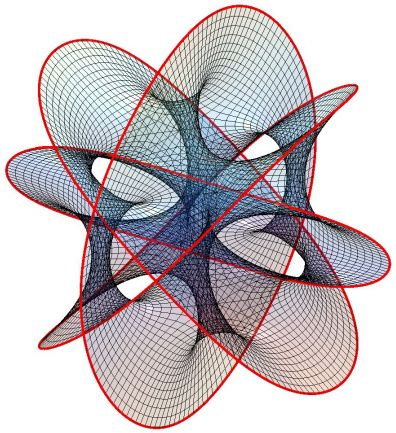
  where J_q is Jacobian of function $q: T_xM \rightarrow R^{n+k}$
- Different metric for every point on the manifold - tensor field



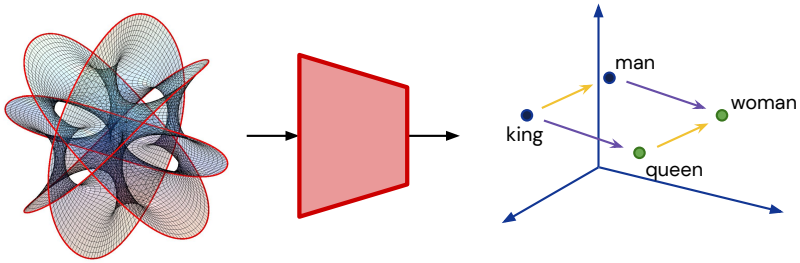
n-dimensional manifold **M**

embedded in (n+k)-dimensional Euclidean space

Two options for building models

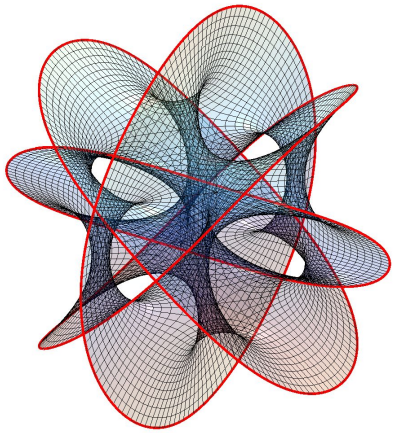


- Stay in original data manifold
- Learn appropriate atlas and metric

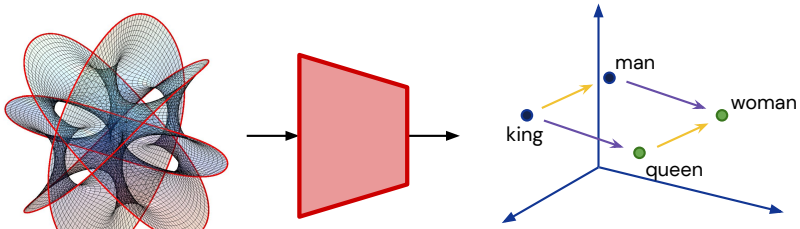


- Move away from original data manifold
- Assume (Euclidean) manifold with simple atlas and metric
- Learn projection to such manifold

Two options for building models



- Stay in original data manifold
- Learn appropriate atlas and metric



- Move away from original data manifold
- Assume (Euclidean) manifold with simple atlas and metric
- Learn projection to such manifold

Spectral/kernel methods

Want to learn more?

Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Schölkopf, Neural Computation 1998

A Global Geometric Framework for Nonlinear Dimensionality Reduction, Tennenbaum et al, Science 2000

Visualizing Data using t-SNE, van der Maaten & Hinton, JMLR 2008

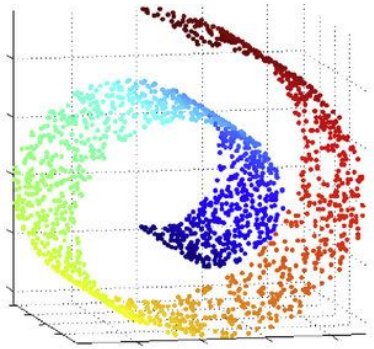
- Calculate data similarity
  - Covariance (PCA)
  - Euclidean distance (MDS, Isomap, LLE, Laplacian Eigenmap)
  - Stochastic neighbor embedding (t-SNE, UMAP)

- Perform spectral decomposition (PCA, kernel PCA, MDS, Isomap, LLE, Laplacian Eigenmap)

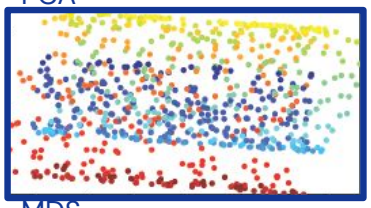
  and/or

- Optimise soft loss (MDS, Isomap, t-SNE, UMAP)

Original data

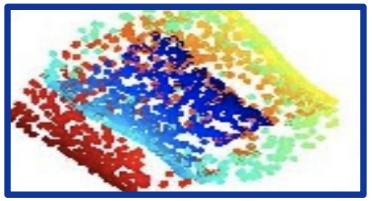


PCA



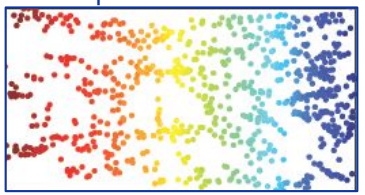
kernel PCA

MDS

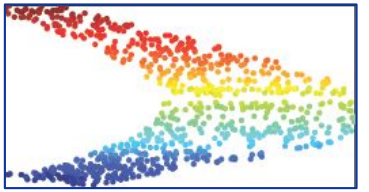


t-SNE

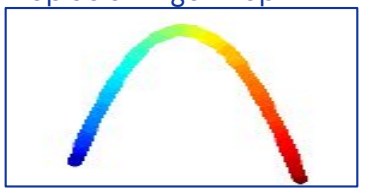
Isomap



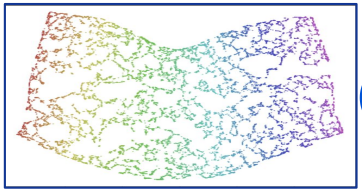
LLE



Laplacian Eigenmap



UMAP



Spectral/kernel methods

- Calculate data similarity
  - Covariance (PCA)
  - Euclidean distance (MDS, Isomap, LLE, Laplacian Eigenmap)
  - Stochastic neighbor embedding (t-SNE, UMAP)
  - Kernels (kernel PCA)
- (optional) Build neighbour graph (Isomap, LLE, t-SNE, UMAP, Laplacian Eigenmap)
- Perform spectral decomposition (PCA, kernel PCA, MDS, Isomap, LLE, Laplacian Eigenmap)

  and/or
- Optimise soft loss (MDS, Isomap, t-SNE, UMAP)

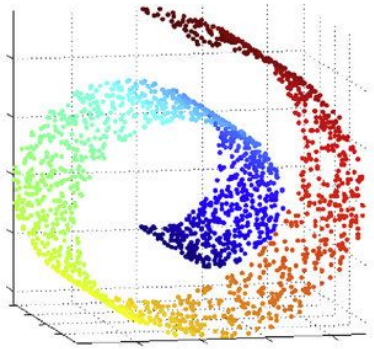
Want to learn more?

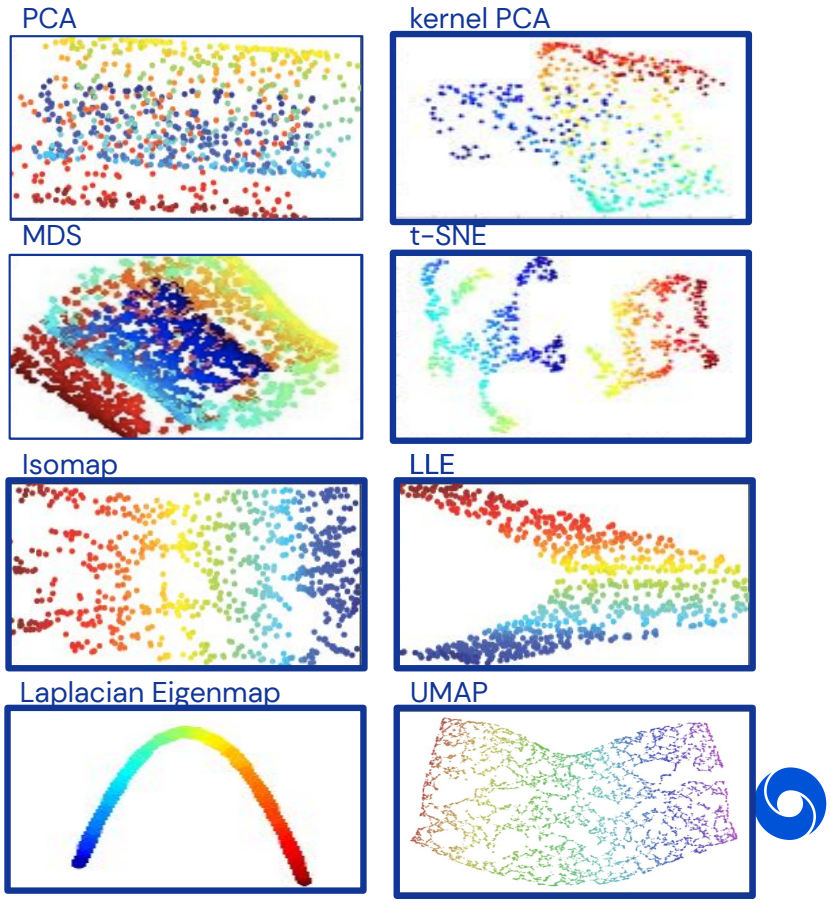
Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Schölkopf, Neural Computation 1998

A Global Geometric Framework for Nonlinear Dimensionality Reduction, Tennenbaum et al, Science 2000

Visualizing Data using t-SNE, van der Maaten & Hinton, JMLR 2008

Original data

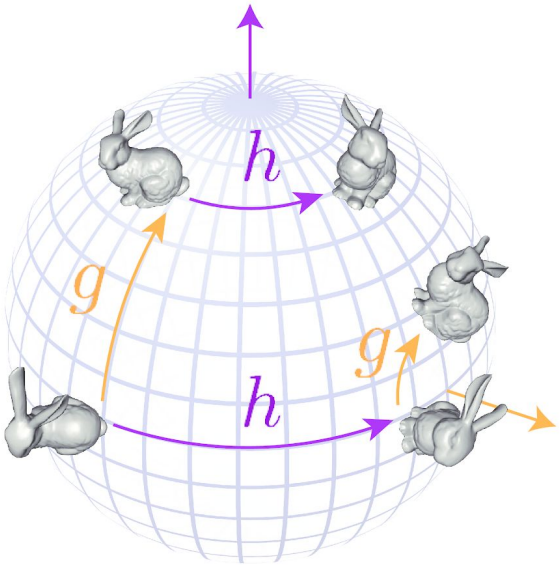




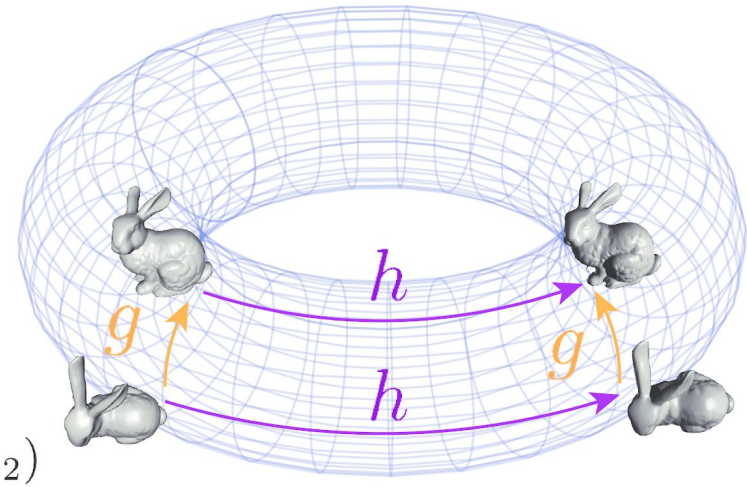
Geometric Manifold Component Estimator (GEOMANCER)

Want to learn more?

Disentangling by Subspace Diffusion, Pfau et al, arXiv 2020



$$g \cdot h \neq h \cdot g$$

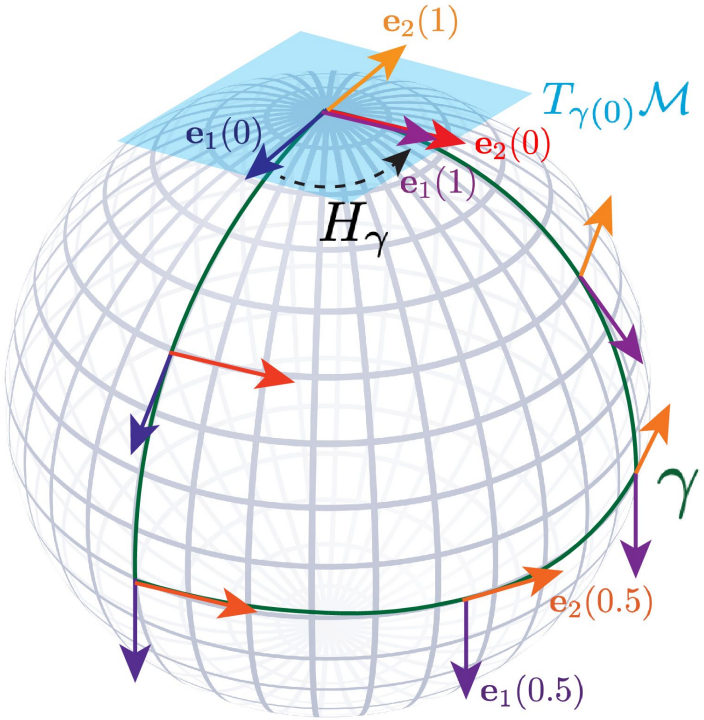


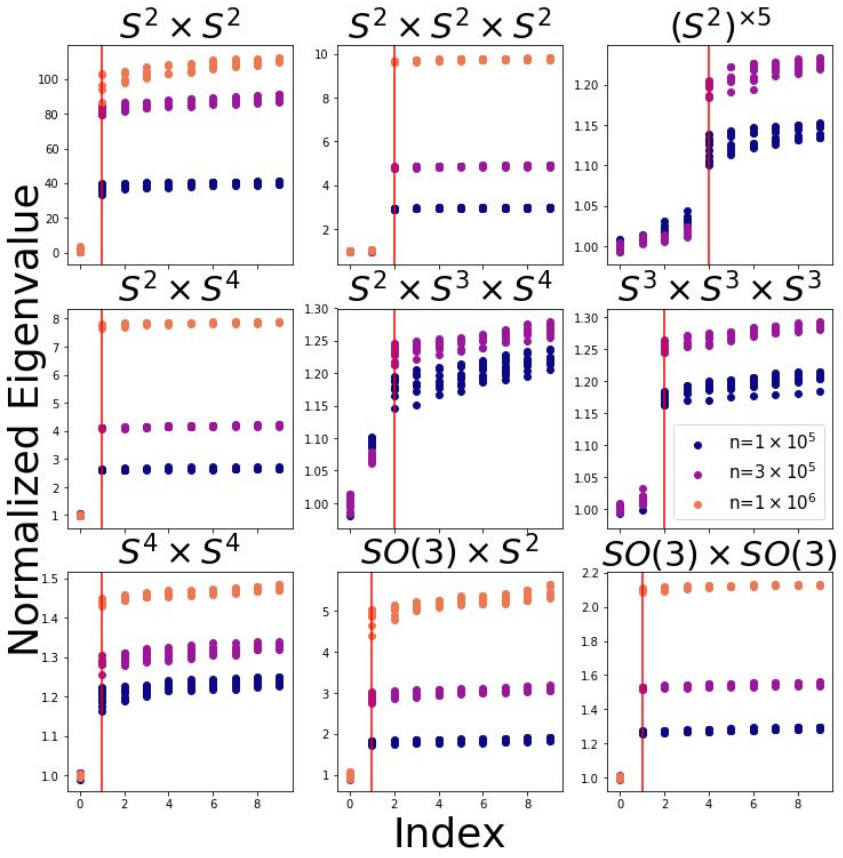
$$g = (g_1, e_2)$$
$$h = (e_1, h_2)$$
$$g \cdot h = h \cdot g = (g_1, h_2)$$

Geometric Manifold Component Estimator (GEOMANCER)

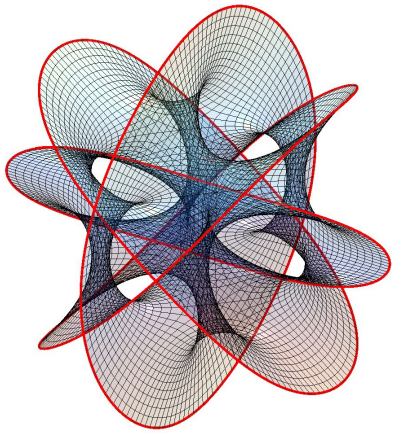
Want to learn more?

Disentangling by Subspace Diffusion, Pfau et al, arXiv 2020

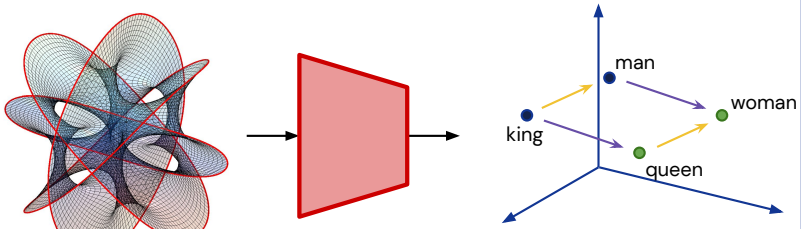




Two options for building models



- Stay in original data manifold
- Learn appropriate atlas and metric



- Move away from original data manifold
- Assume (Euclidean) manifold with simple atlas and metric
- Learn projection to such manifold

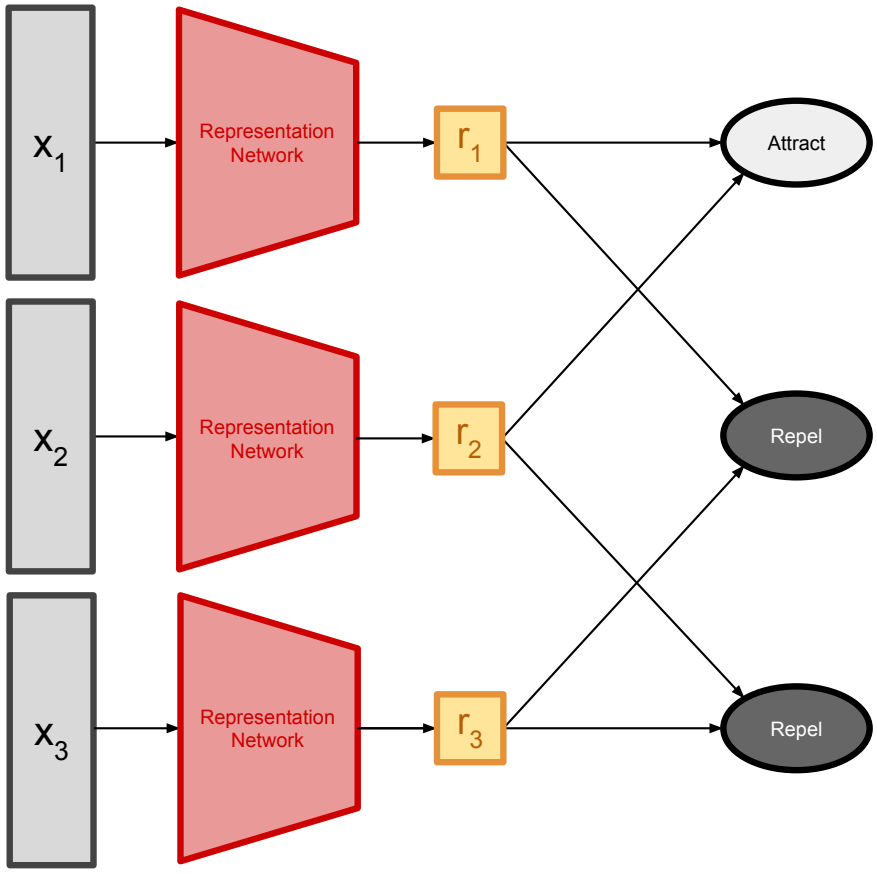
Contrastive learning / energy models / metric learning

Want to learn more?

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, Grill et al, arxiv 2020

Data-Efficient Image Recognition with Contrastive Predictive Coding, Hénaff et al, ICML 2020

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML 2020



Positive pair

Negative pair

Negative pair

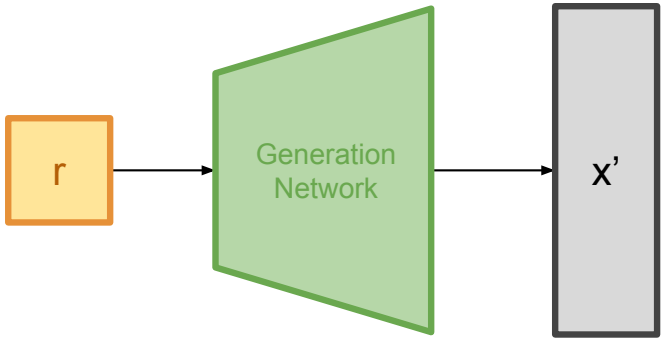
$$\mathcal{L}_{\text{SimCLR}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

$$\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$$

$$\mathcal{L}_\theta^{\texttt{BYOL}} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2$$

$$\mathcal{L}_{\text{CPC}} = -\sum_{i,j,k} \log \frac{\exp(\hat{\boldsymbol{z}}_{i+k,j}^T \boldsymbol{z}_{i+k,j})}{\exp(\hat{\boldsymbol{z}}_{i+k,j}^T \boldsymbol{z}_{i+k,j}) + \sum_l \exp(\hat{\boldsymbol{z}}_{i+k,j}^T \boldsymbol{z}_l)}$$

Adversarial learning

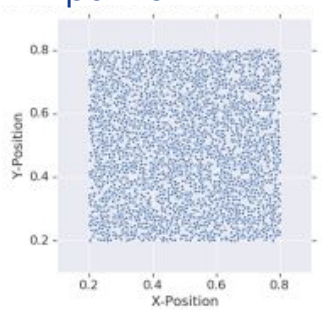


Variational Autoencoder learning

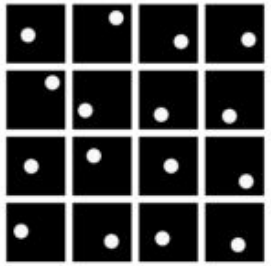
Want to learn more?

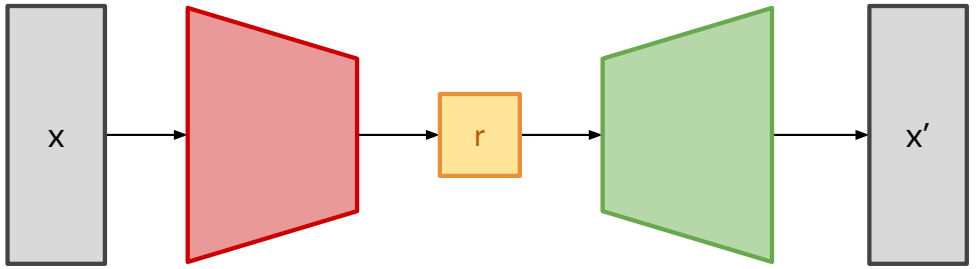
Spatial Broadcast Decoder: A Simple Architecture for Learning Disentangled Representations in VAEs, Watters et al, ICLR workshop 2018

Training data points

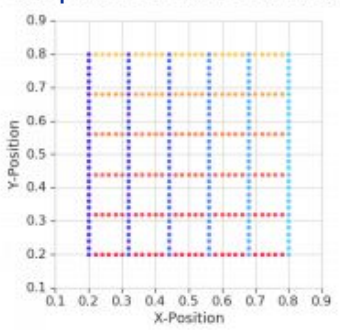


Dataset samples



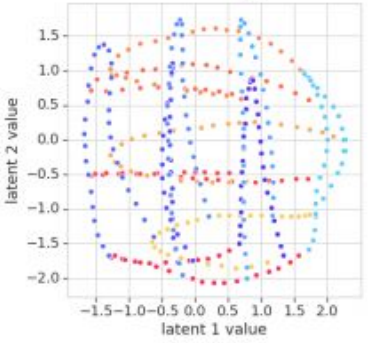


Test data points

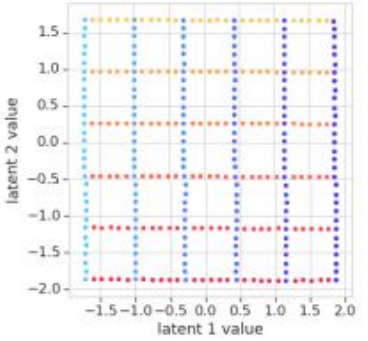


Pretrained encoder

VAE



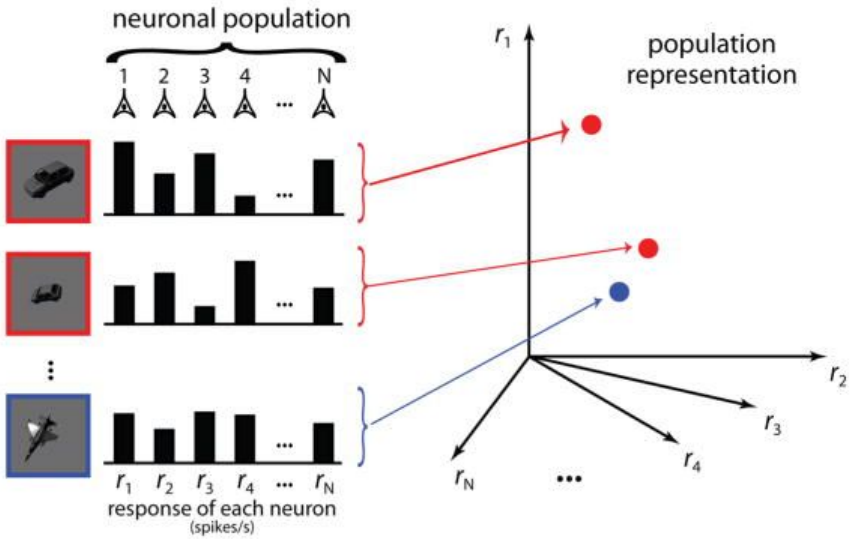
Factor VAE w\ BD



Untangling representations

Want to learn more?

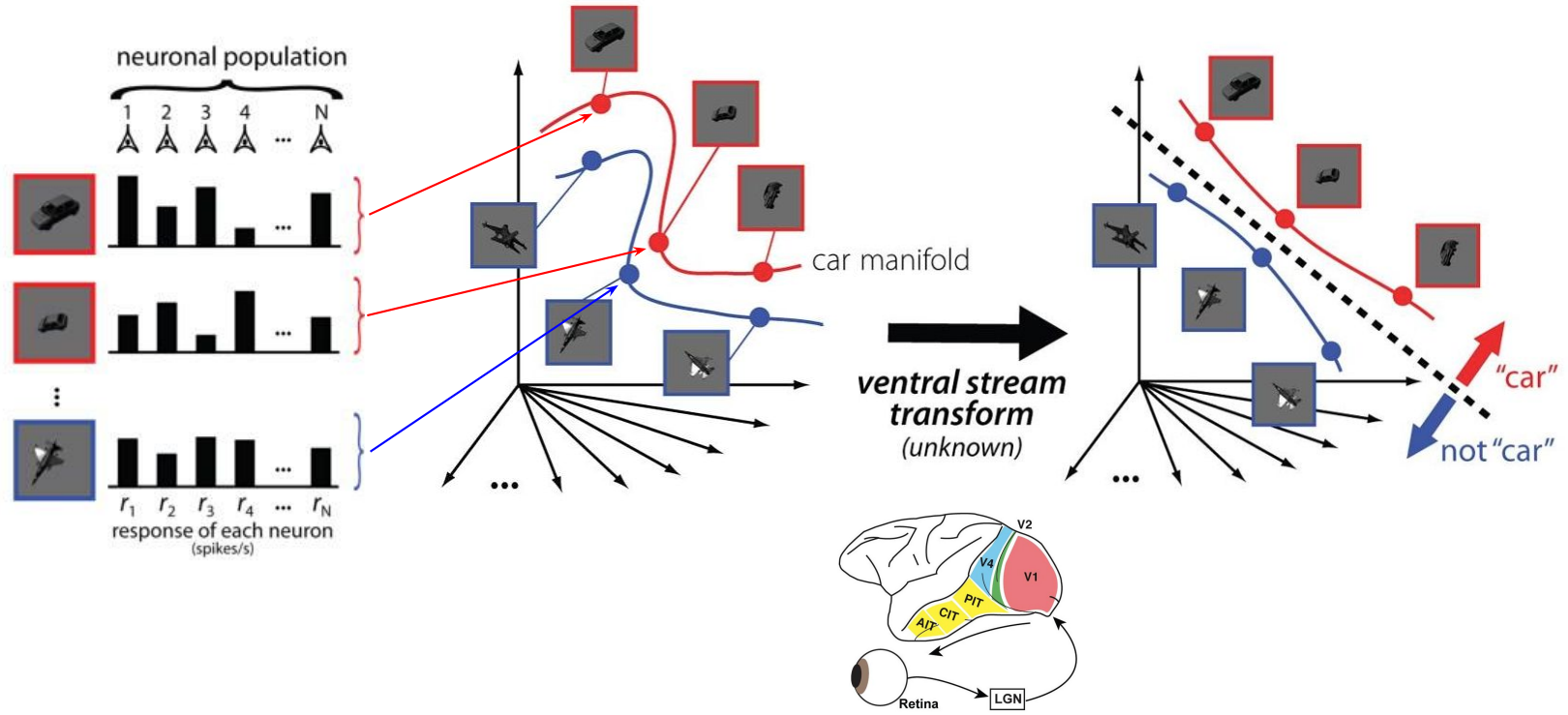
How Does the Brain Solve Visual Object Recognition?, DiCarlo et al, Neuron 2012



Untangling representations

Want to learn more?

How Does the Brain Solve Visual Object Recognition?, DiCarlo et al, Neuron 2012



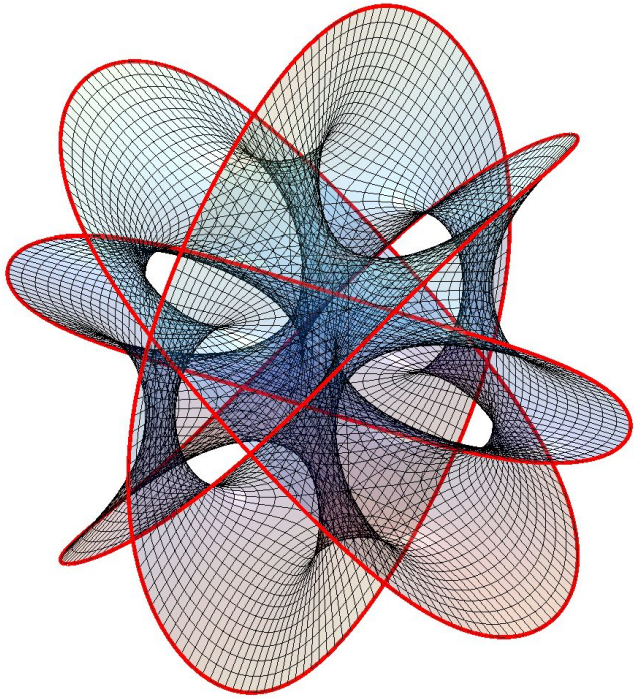
Summary

Thinking about the topology of the original **data manifold**,

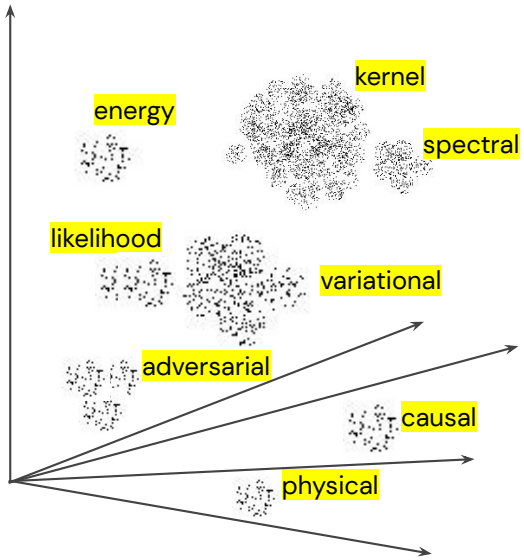
and the explicit or implicit **restrictions** on the topology of the learnt **representation manifold**

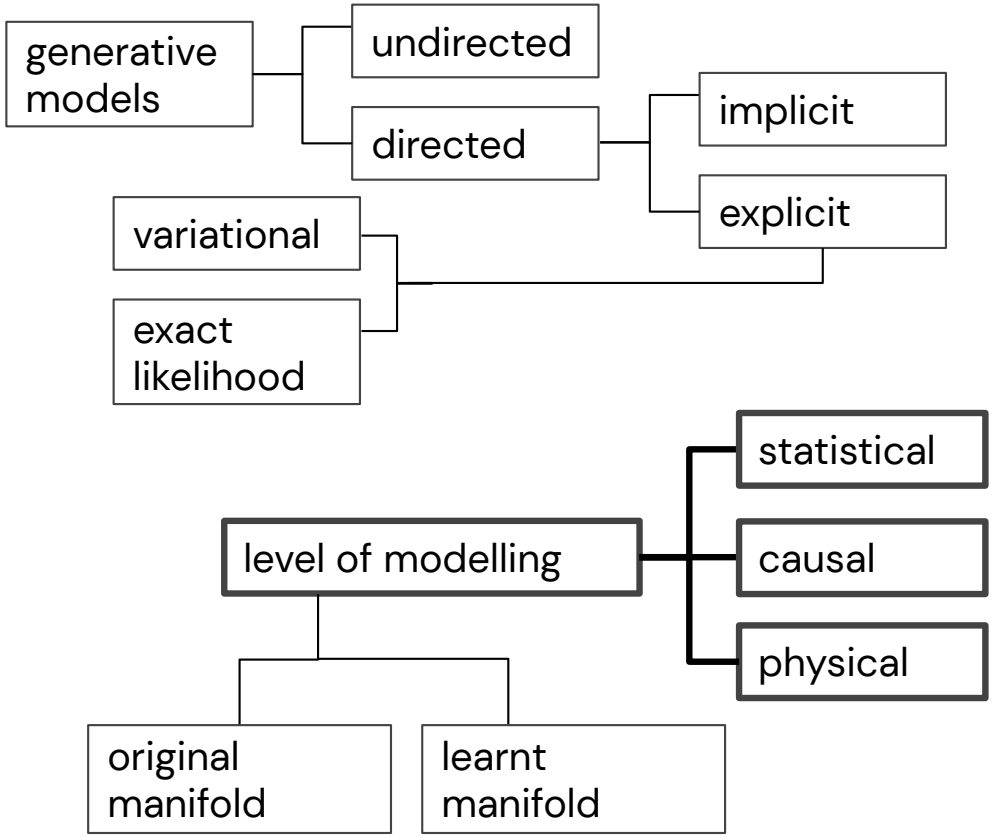
may help **troubleshoot problems** with the existing representation learning approaches

and help **develop better methods**.



Mapping out the landscape





Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- IID/non-IID (interventional)
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

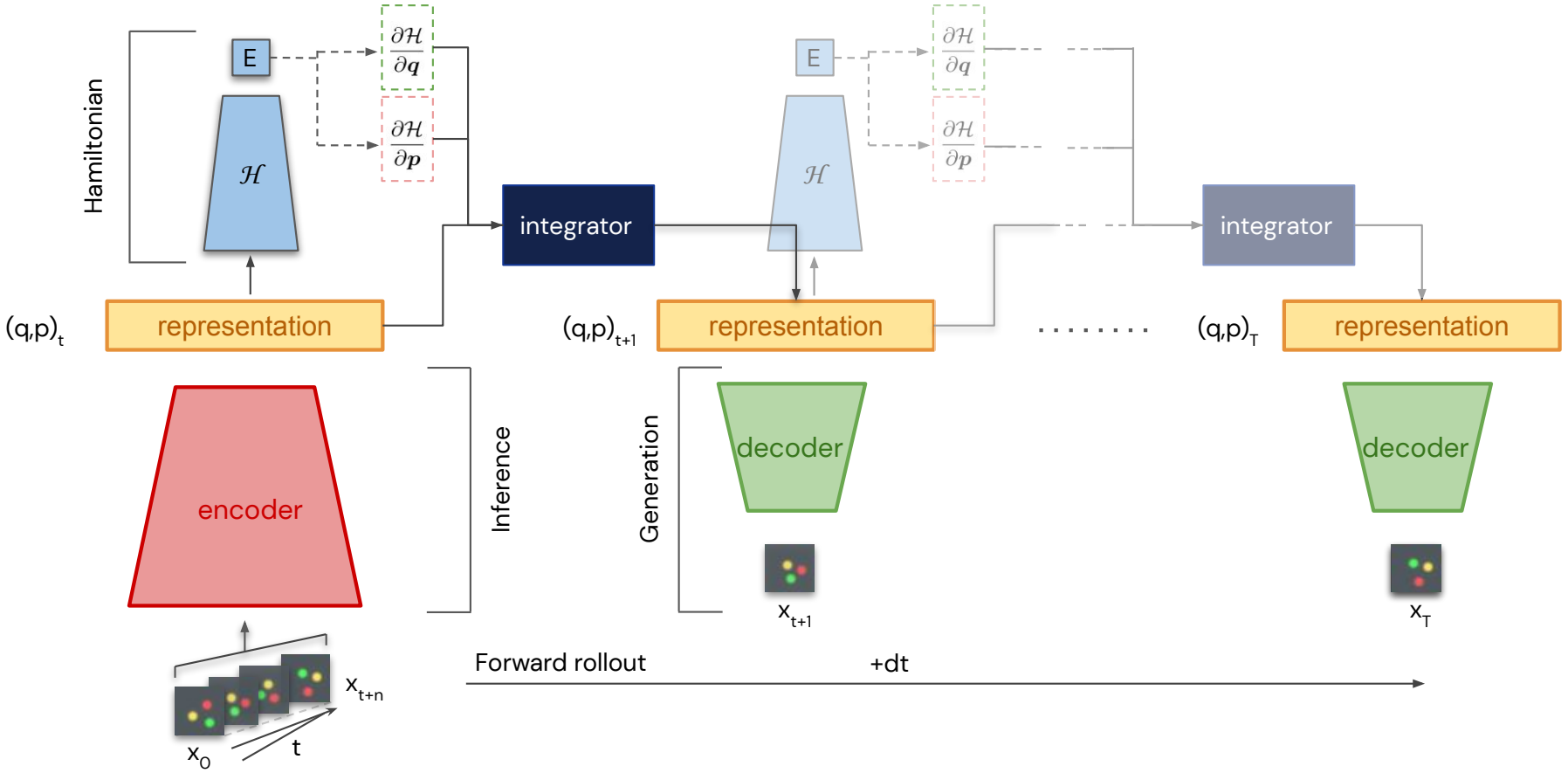
Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

Hamiltonian Generative Network (HGN)

Want to learn more?

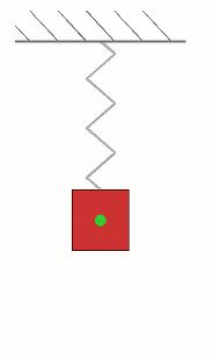
Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020

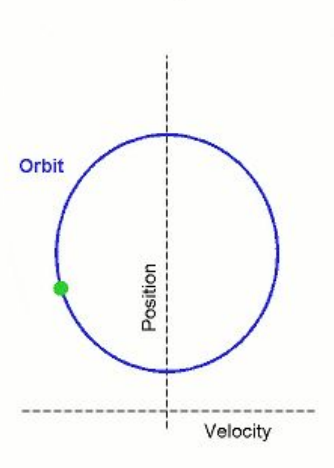


What is a Hamiltonian?

Want to learn more?

Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020





q: position

p: momentum

phase space

Hamiltonian: $H(q,p) = \frac{1}{2}kq^2 + \frac{p^2}{2m}$

Time evolution: $\frac{dq}{dt} = \frac{\partial H}{\partial p}$ $\frac{dp}{dt} = -\frac{\partial H}{\partial q}$

E.g. with Euler integrator:

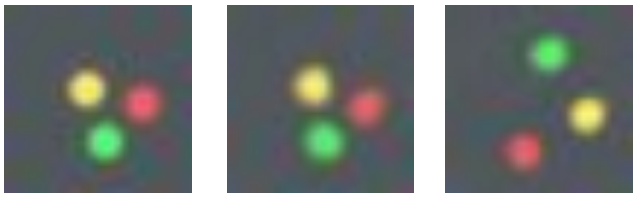
$q_{t+1} = q_t + dt \frac{\partial H}{\partial p}$ $p_{t+1} = p_t - dt \frac{\partial H}{\partial q}$

Power of Physical modeling

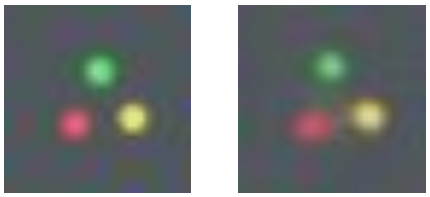
Want to learn more?

Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020

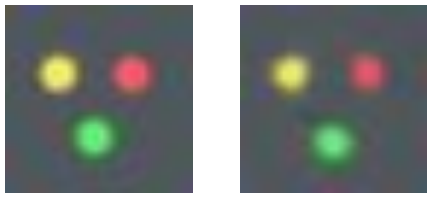
Original Reconstructed Reversed



Original 2x slower



Original 2x faster



Samples

Sample 1 Sample 2 Sample 3 Sample 4 Sample 5 Sample 6

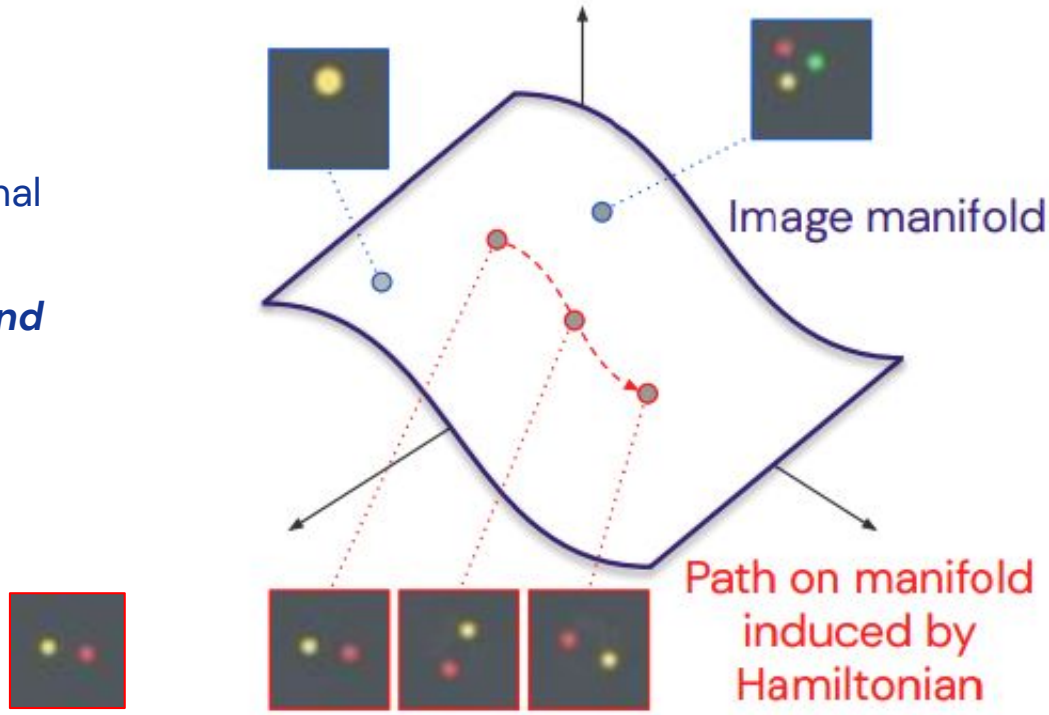
Hamiltonian Manifold Hypothesis

Want to learn more?

Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020

Natural images lie on a low dimensional manifold in pixel space and

natural image sequences correspond to motion governed by abstract Hamiltonian dynamics.



Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- IID/non-IID (interventional)
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

Statistical perspective of disentangling

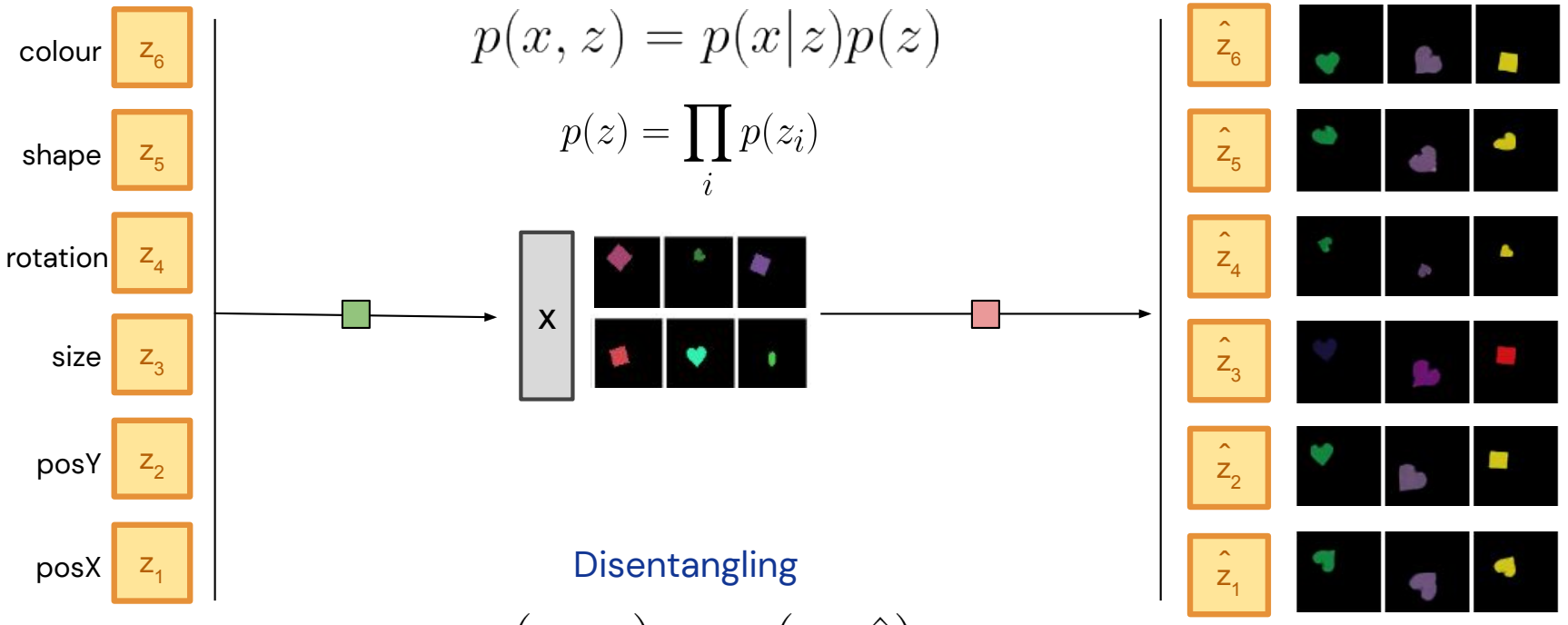
Want to learn more?

Deep Learning of Representations: Looking Forward, Bengio, SLSP 2013

Generative model

$$p(x, z) = p(x|z)p(z)$$

$$p(z) = \prod_i p(z_i)$$



Disentangling

$$p(x, z) = p(x, \hat{z})$$

GIFs adapted from Chris Burgess

Statistical perspective of disentangling

Want to learn more?

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, Locatello et al, ICML 2019

Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathbf{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

- Disentangled representations are **non-identifiable** in naive unsupervised setting
- Inductive **biases in models and/or data** make unsupervised disentangling work in practice (e.g. β-VAE, FactorVAE, TC-VAE etc)

Statistical perspective of disentangling

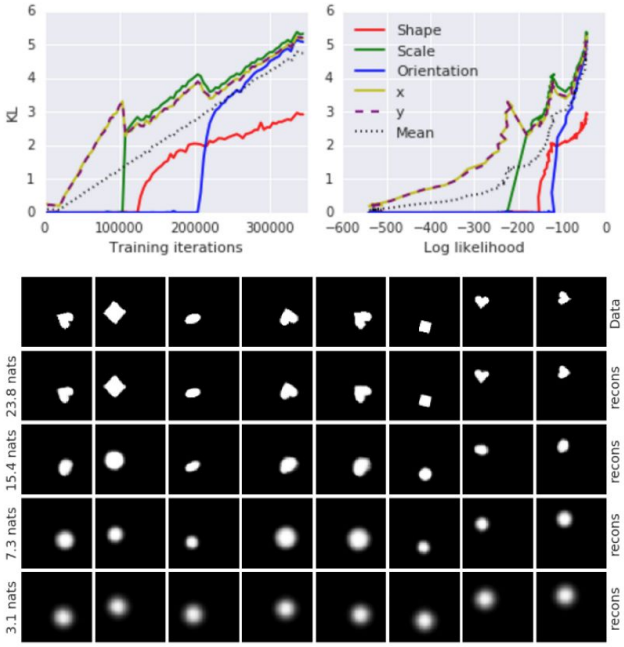
Want to learn more?

Understanding Disentangling in β-VAE. Burgess et al, arxiv 2018

Variational Autoencoders Pursue PCA Directions (by Accident). Rolinek, Zietlow et al, CVPR 2019

Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathbf{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

- Disentangled representations are **non-identifiable** in naive unsupervised setting
- Inductive **biases in models and/or data** make unsupervised disentangling work in practice (e.g. β-VAE, FactorVAE, TC-VAE etc)



Optimizing the stochastic part of the reconstruction loss promotes local orthogonality of the decoder.

Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- IID/non-IID (interventional)
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

Structural causal models (SCMs)

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

Given set of observables $x_1, \ldots, x_n$ (random variables, vertices of a DAG), each can be expressed as:

$$x_i := f_i(\mathbf{PA}_i, U_i)$$

$\mathbf{PA}_i$ - parents

f_i - deterministic function

U_i - stochastic unexplained variable, where $U_1, \ldots, U_n$ are jointly independent

Given SCM can express conditional probabilities $p(x_i|\mathbf{PA}_i)$

Independent causal mechanisms

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

Independent Causal Mechanisms (ICM) Principle. *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

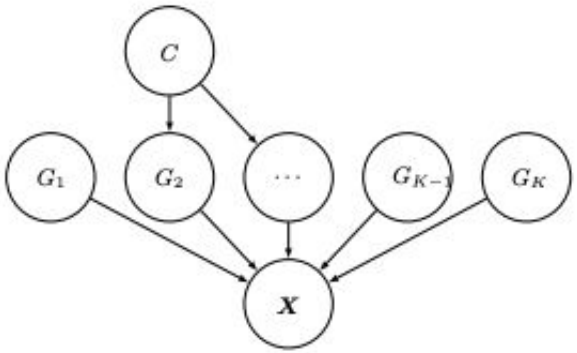
In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

- Changing (or intervening upon) one mechanism $p(x_i|\mathbf{PA}_i)$ does not change other mechanisms $p(x_j|\mathbf{PA}_j),\ j \neq i$
- Knowing about other mechanisms $p(x_j|\mathbf{PA}_j),\ j \neq i$ does not inform about mechanism $p(x_i|\mathbf{PA}_i)$

Causal view of disentangling

Want to learn more?

Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness, Suter et al, ICML 2019



Definition 1 (Disentangled Causal Process). *Consider a causal model for $\boldsymbol{X}$ with generative factors $\boldsymbol{G}$, described by the mechanisms $p(\boldsymbol{x}|\boldsymbol{g})$, where $\boldsymbol{G}$ could generally be influenced by L confounders $\boldsymbol{C} = (C_1, \ldots, C_L)$. This causal model for $\boldsymbol{X}$ is called disentangled if and only if it can be described by a structural causal model (SCM) (Pearl, 2009) of the form*

$$\boldsymbol{C} \leftarrow \boldsymbol{N}_c$$
$$G_i \leftarrow f_i(\boldsymbol{PA}_i^C, N_i), \quad \boldsymbol{PA}_i^C \subset \{C_1, \ldots, C_L\},\ i = 1, \ldots, K$$
$$\boldsymbol{X} \leftarrow g(\boldsymbol{G}, N_x)$$

with functions f_i, g and jointly independent noise variables $\boldsymbol{N}_c, N_1, \ldots, N_K, N_x$. Note that $\forall i \neq j\ G_i \not\to G_j$.

Proposition 1 (Properties of a Disentangled Causal Process). *A disentangled causal process as introduced in Definition 1 fulfills the following properties:*

(a) $p(\boldsymbol{x}|\boldsymbol{g})$ *describes a causal mechanism invariant to changes in the distributions* $p(g_i)$.

(b) In general, the latent causes can be dependent

$$G_i \not\perp\!\!\!\perp G_j, i \neq j.$$

Only if we condition on the confounders in the data generation they are independent

$$G_i \perp\!\!\!\perp G_j | \boldsymbol{C} \quad \forall i \neq j.$$

(c) Knowing what observation of $\boldsymbol{X}$ we obtained renders the different latent causes dependent, i.e.,

$$G_i \not\perp\!\!\!\perp G_j | \boldsymbol{X}.$$

(d) The latent factors $\boldsymbol{G}$ already contain all information about confounders $\boldsymbol{C}$ that is relevant for $\boldsymbol{X}$, i.e.,

$$I(\boldsymbol{X}; \boldsymbol{G}) = I(\boldsymbol{X}; (\boldsymbol{G}, C)) \geq I(\boldsymbol{X}; C)$$

where I denotes the mutual information.

(e) There is no total causal effect from G_j to G_i for $j \neq i$; i.e., intervening on G_j does not change G_i, i.e,

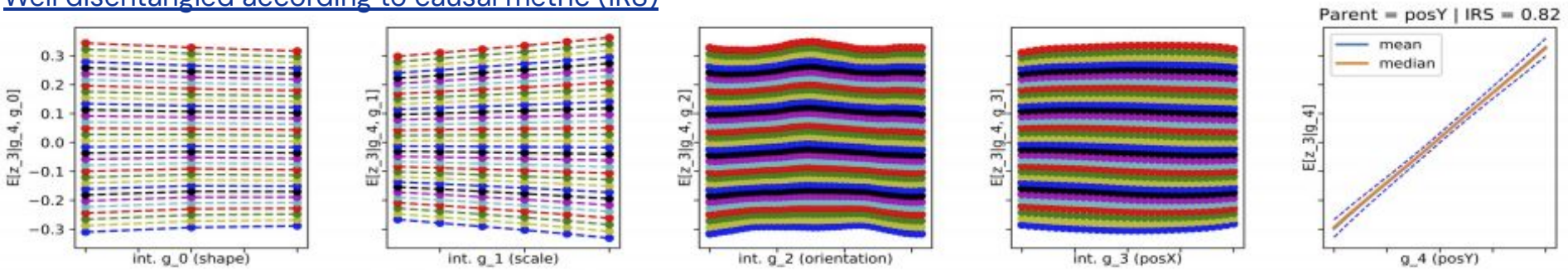
$$\forall g_j^{\triangle} \quad p(g_i | do(G_j \leftarrow g_j^{\triangle})) = p(g_i) \quad \left(\neq p(g_i | g_j^{\triangle})\right)$$

Measuring disentangling through causality

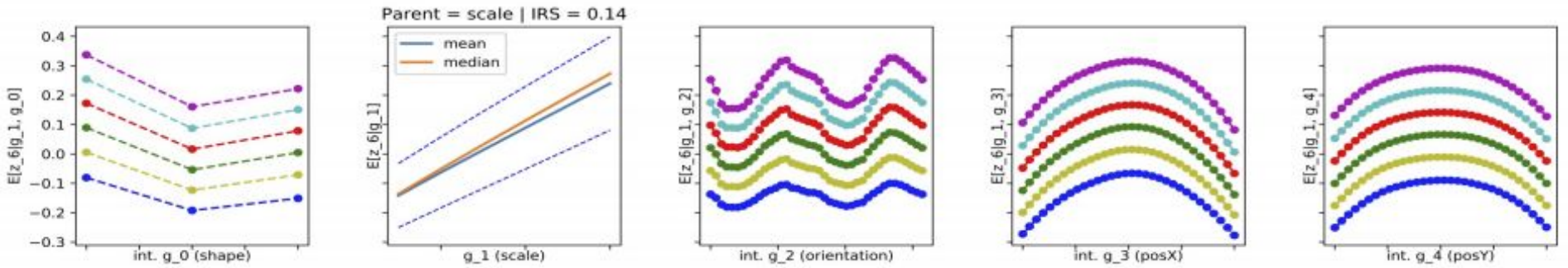
Want to learn more?

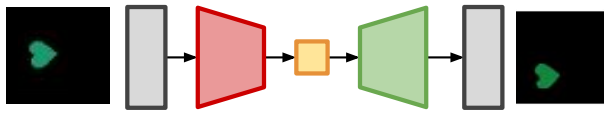
Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness, Suter et al, ICML 2019

Well disentangled according to causal metric (IRS)



Badly disentangled according to causal metric (IRS)





Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- Non-IID
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

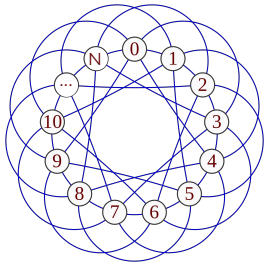
Physical definition of disentangling

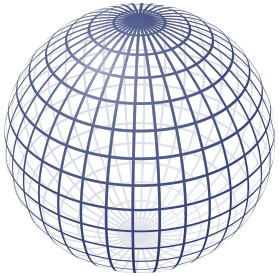
Want to learn more?

Towards a Definition of Disentangled Representations, Higgins, Amos et al, ICML Workshop on Theoretical Physics for Deep Learning 2019

A vector representation is called a ***disentangled representation*** *with respect to a particular decomposition of a symmetry group into subgroups, if it decomposes into independent subspaces, where each subspace is affected by the action of a single subgroup, and the actions of all other subgroups leave the subspace unaffected.*

Primer on group theory: examples





$$R_z(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Cyclic group - C_N

Group generated by a single element g. Every element of the group may be obtained by repeatedly applying the group operation to g or its inverse.

Rotational symmetries of a polygon forms a finite cyclic group.

Describes integers modulo N.

3D rotation group - SO(3)

Group of all rotations about the origin of 3D Euclidean space R^3 under the operation of composition.

Group of all orthogonal 3x3 matrices with determinant 1.

Describes rotational symmetries of 3D objects, orientations.

Primer on group theory: group and group action

A **group** is a set G with an operator $\circ: G \times G \to G$:

$$G = \{ \{e, g_1, g_2\}, \circ \}$$

A group has four properties:

1) Associativity $\forall\ x,y,z \in G : x \circ (y \circ z) = (x \circ y) \circ z$
2) Identity $\exists\ e \in G\ \forall\ x \in G : e \circ x = x \circ e = x$
3) Inverse $\forall\ x \in G\ \exists\ x^{-1} \in G : x \circ x^{-1} = x^{-1} \circ x = e$

Primer on group theory: group action

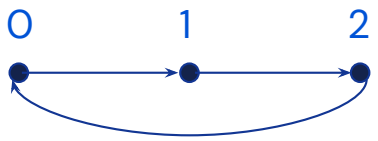
Group **action** on (set) W:

$\cdot : G \times W \rightarrow W$

should satisfy:

1) Identity $ew = w \quad \forall \; w \in W$
2) Associativity $(gh)w = g(hw) \quad \forall \; g, h \in G, w \in W$

$W = \{ [0], [1], [2] \}$

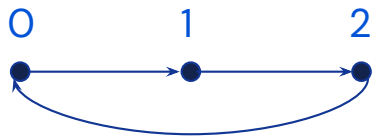


	e	$\mathbf{g_1}$	$\mathbf{g_2}$
e	e	g_1	g_2
$\mathbf{g_1}$	g_1	g_2	e
$\mathbf{g_2}$	g_2	e	g_1

Primer on group theory: subgroup

$G = \{ \{e, g_1, g_2\}, \circ \}$

$W = \{ [0], [1], [2] \}$



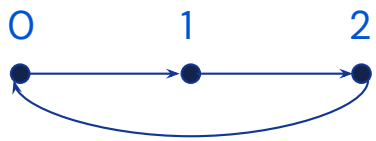
	e	$\mathbf{g_1}$	$\mathbf{g_2}$
e	e	g_1	g_2
$\mathbf{g_1}$	g_1	g_2	e
$\mathbf{g_2}$	g_2	e	g_1

$G = C_3$

Primer on group theory: subgroup

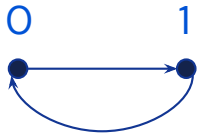
$G = \{ \{e, g_1, g_2\}, \circ \}$

$W = \{ [0], [1], [2] \}$



H	**e**	**g_1**	g_2
e	e	g_1	g_2
g_1	g_1	e	e
g_2	g_2	e	g_1

$G = C_3$ $H = C_2$



H is a **subgroup** of G, if it is a subset of G, that is closed under $\circ$ operator and inverses:

$x,y \in H \Rightarrow x \circ y \in H \wedge x^{-1} \in H$

Primer on group theory: direct product

Given two groups G and H, can construct a new group K = G x H (**direct product**):

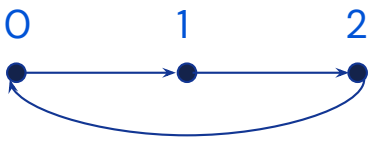
1) Underlying set is a cartesian product G x H, i.e. ordered pairs (g, h) where $g \in G$ and $h \in H$

2) Group operator • defined component wise, i.e. $(g_1, h_1) \bullet (g_2, h_2) = (g_1 \circ g_2, h_1 \odot h_2)$

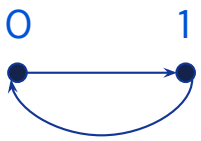
$K = G \times H$

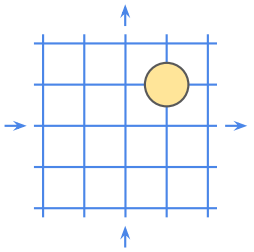
G \ H	$e_g e_h$	$g_1 e_h$	$g_2 e_h$	$e_g h_1$	$g_1 h_1$	$g_2 h_1$
$e_g e_h$	$e_g e_h$	$g_1 e_h$	$g_2 e_h$	$e_g h_1$	...	...
$g_1 e_h$	$g_1 e_h$	$g_2 e_h$	$e_g e_h$	...	...	...
$g_2 e_h$	$g_2 e_h$	$e_g e_h$	$g_1 e_h$	...	...	...
$e_g h_1$	$e_g h_1$	...	...	$e_g e_h$	...	...
$g_1 h_1$	...	...	...	...	...	...
$g_2 h_1$	...	...	...	...	...	...

$G = \{ \{e_g, g_1, g_2\}, \circ \}$



$H = \{ \{e_h, h_1\}, \odot \}$





Position y

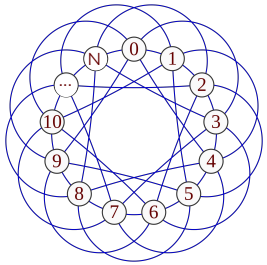
Position x

Colour

$G_x = C_N$

$G_y = C_N$

$G_c = C_N$



Want to learn more?

Towards a Definition of Disentangled Representations, Higgins, Amos et al, ICML Workshop on Theoretical Physics for Deep Learning 2019

Symmetry group $G = G_x \times G_y \times G_c$

W

(**x=5**, y=6, c=yellow)

$\cdot : G \times W \to W$

b

O

h

f

Z

$[z_x = \mathbf{0.9}, z_y = 0.3, z_c = 0.1]$

$* : G \times Z \to Z$

Want to find an **equivariant map** f, s.t.:

$g \cdot f(w) = f(g \cdot w) \quad \forall \; g \in G, w \in W$

W

(**x=3**, y=6, c=yellow)

b

O

h

f

Z

$[z_x = \mathbf{0.3}, z_y = 0.3, z_c = 0.1]$

Summary

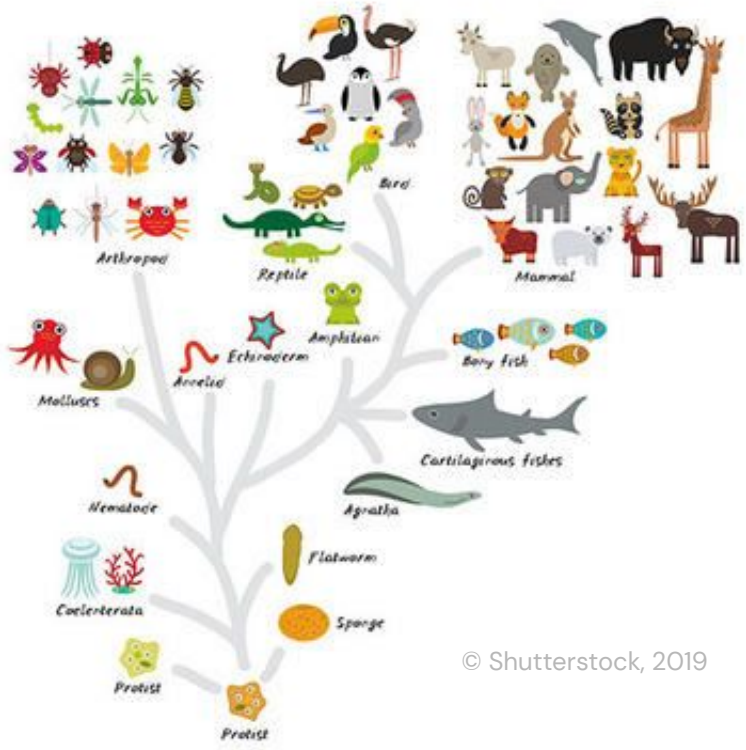
The diversity of models for unsupervised representation learning appears large - the “model zoo”

The large space of models can be understood and navigated using a relatively simple theoretical taxonomy:

1. How is the **data density** modelled?
2. What properties of the **manifold** are modelled?
3. What **level of modeling detail** is expected?

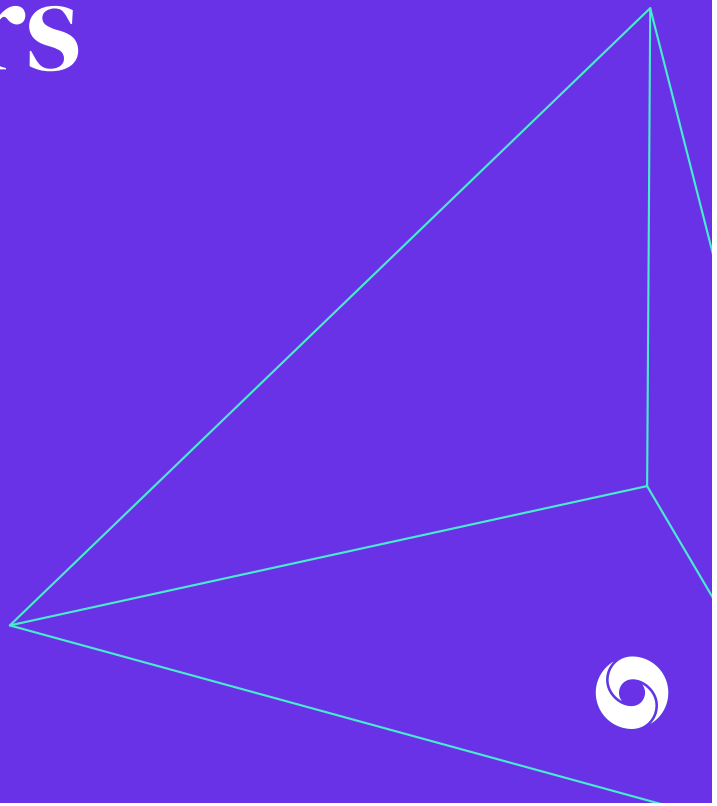
Thinking about these questions may help understand:

1. the **tradeoffs** of different implementational choices
2. **troubleshoot** failures
3. possible directions of model **improvement**

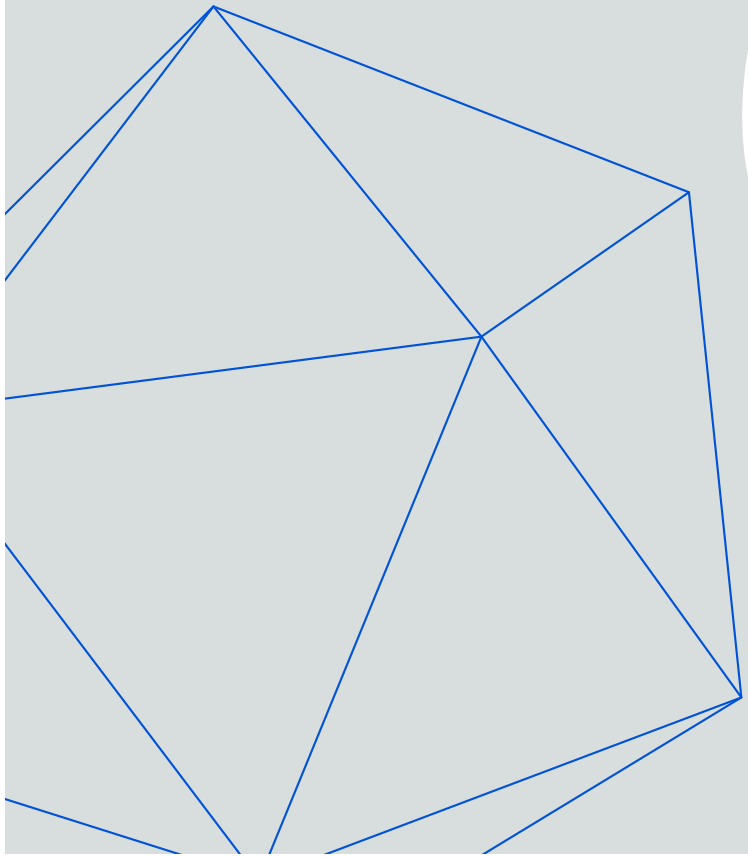


DeepMind

4 Frontiers



DeepMind



Multi-disciplinary perspective

Sources of inspiration

linguistics

physics

machine learning

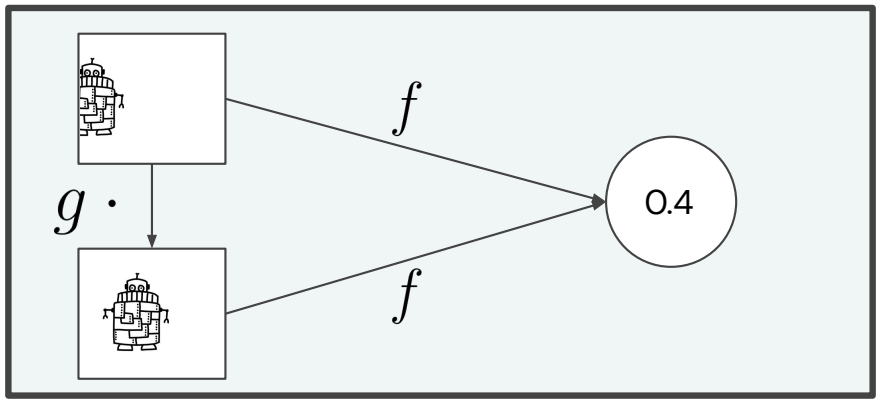
neuroscience

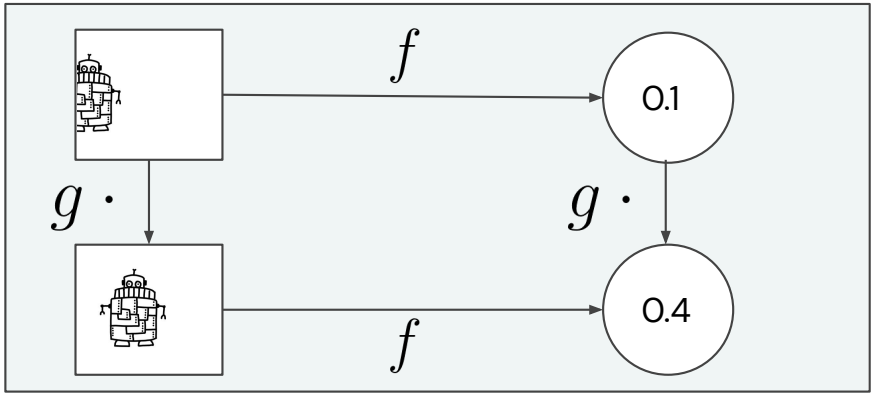
mathematics

Invariance vs equivariance

Want to learn more?

Backpropagation Applied to Handwritten Zip Code Recognition, LeCun et al, Neural Computation 1989





Invariance

- representation remains unchanged when a certain type of transformation is applied to the input

$$f(g \cdot x) = f(x)$$

Equivariance

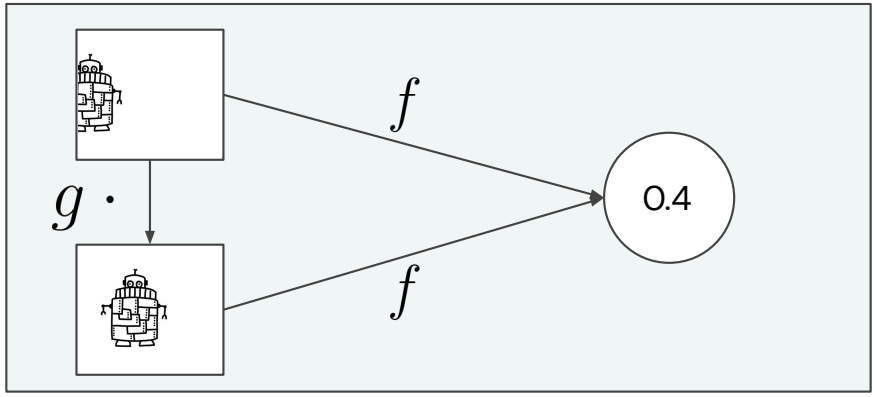
- representation reflects the transformation applied to the input

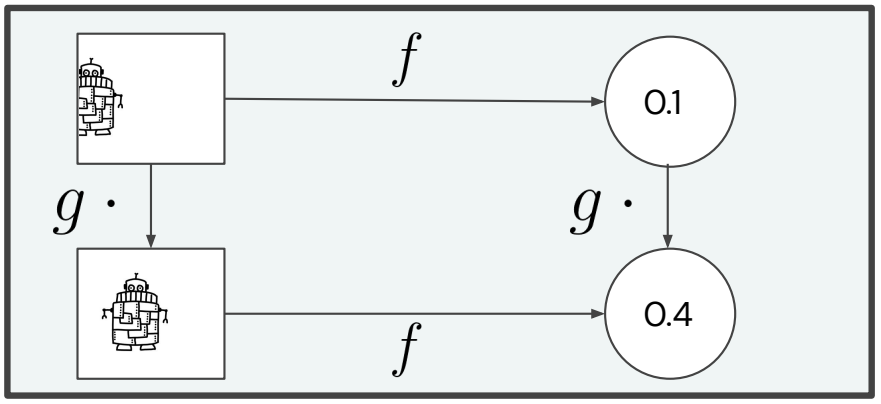
$$f(g \cdot x) = g \cdot f(x)$$

Invariance vs equivariance

Want to learn more?

A General Theory of Equivariant CNNs on Homogeneous Spaces, Cohen et al, NeurIPS 2019





Invariance

- representation remains unchanged when a certain type of transformation is applied to the input

$$f(g \cdot x) = f(x)$$

Equivariance

- representation reflects the transformation applied to the input

$$f(g \cdot x) = g \cdot f(x)$$

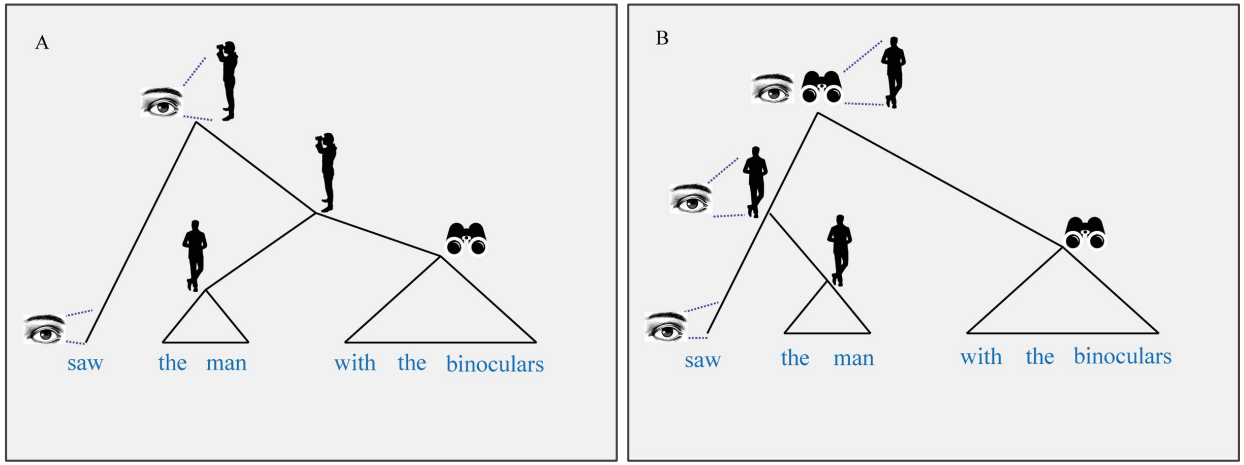
Compositionality

Want to learn more?

SCAN: Learning Hierarchical Compositional Visual Concepts, Higgins et al, ICLR 2018

"the meaning of a complex expression is determined by the meanings of its **constituent expressions** and the **rules** used to **combine** them"

Leads to **open-endedness** -- can construct *arbitrarily large number of meaningful complex expressions* from a *finite number of constituent expressions* and *combination rules*.



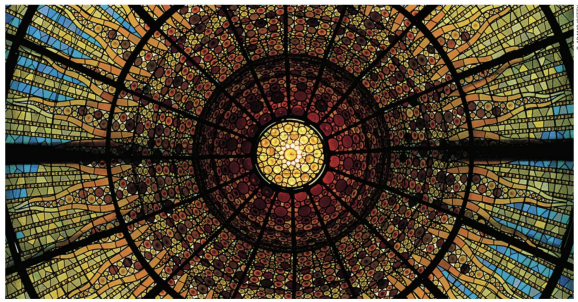
Bolhuis et al, 2018

Symmetry transformations

Want to learn more?

Why symmetry matters, Livio, Nature 2012

COMMENT



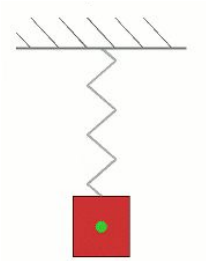
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

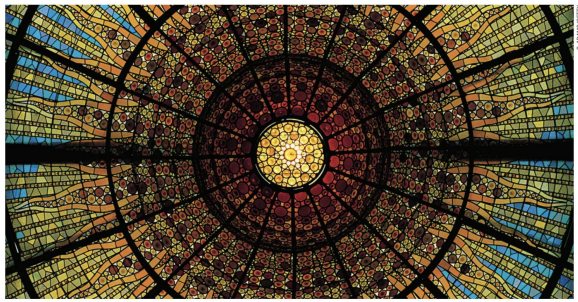
"To a physicist, symmetry is a broader concept than the reflective form of butterfly wings… Symmetry represents those ***stubborn cores that remain unaltered*** *even under transformations that could change them"*

\- Mario Livio, 2012



Symmetry transformations

COMMENT



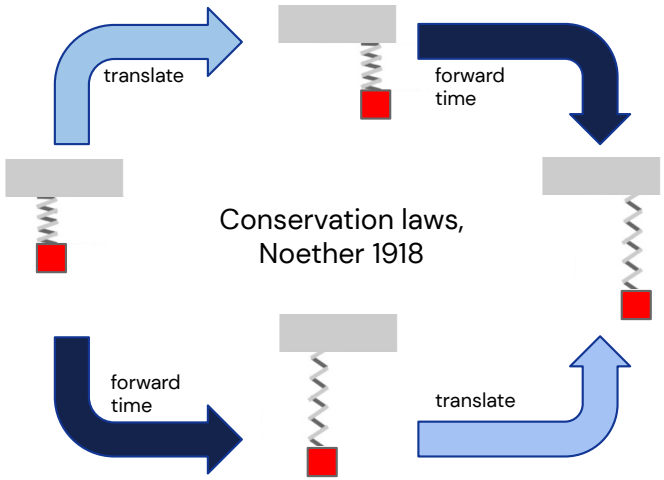
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

*"To a physicist, symmetry is a broader concept than the reflective form of butterfly wings… Symmetry represents those **stubborn cores that remain unaltered** even under transformations that could change them"*

\- Livio, 2012



Conservation laws, Noether 1918

Want to learn more?

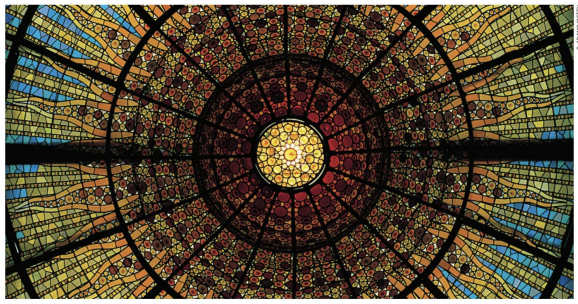
Invariante Variationsprobleme, Noether, Gesellschaft der Wissenschaften zu Göttingen, 1918

Studying symmetries of a system helps:

- Unify existing theories (e.g. electromagnetism)
- Categorise physical objects (e.g. elementary particles)
- Discover new physical objects (e.g. particle Ω- predicted in 1962, discovered in 1964)

Symmetry transformations

COMMENT



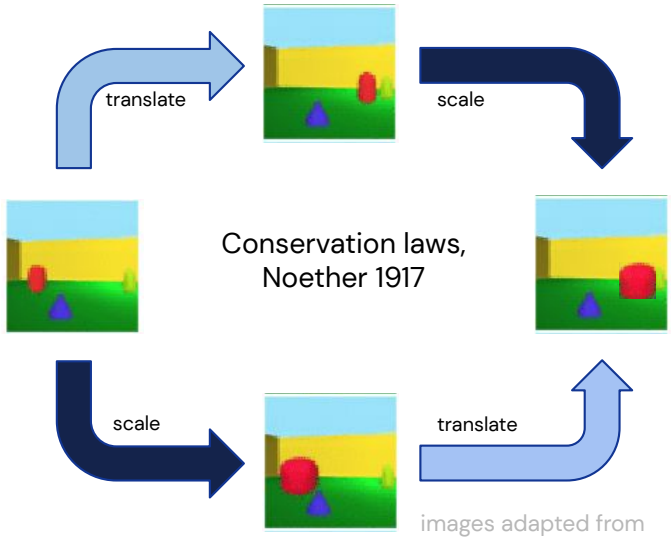
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

*"To a physicist, symmetry is a broader concept than the reflective form of butterfly wings… Symmetry represents those **stubborn cores that remain unaltered** even under transformations that could change them"*

\- Livio, 2012



images adapted from Chris Burgess

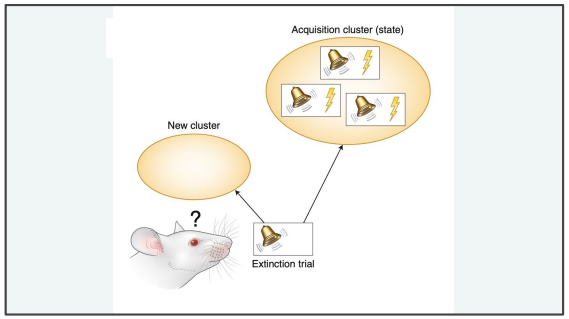
Studying symmetries of a system helps:

- Unify existing theories (e.g. electromagnetism)
- Categorise physical objects (e.g. elementary particles)
- Discover new physical objects (e.g. particle Ω- predicted in 1962, discovered in 1964)

Solving tasks requires...

Want to learn more?

Learning Task-State Representations, Niv, Nature Neuroscience 2019



Attention

Representation should support easy attentional attenuation of aspects not relevant to the task.

Clustering

Experiences should be easily and dynamically clustered together or apart.

Latent states

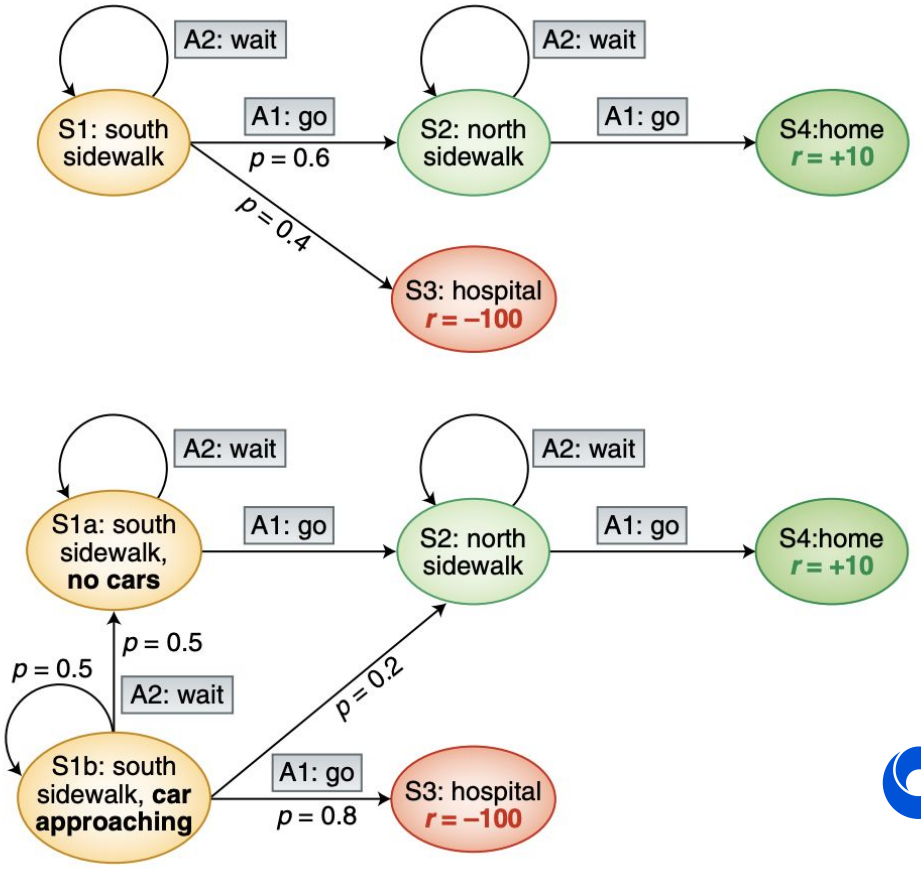
Not all information may be present in perceptual input. Representations should include information about latent aspects of the state too.


How does one cross a street?
Want to learn more?
Learning Task-State Representations, Niv, Nature Neuroscience 2019
BUSES ONLY
4PM-7PM MON THRU FRI
7AM - 7PM MON THRU FRI
EXCEPT DELIVERIES THIS BLOCK
W 25 ST
ONE WAY
PIZZA
BUS ONLY

Alternative representations for the same task

Want to learn more?

Learning Task-State Representations, Niv, Nature Neuroscience 2019



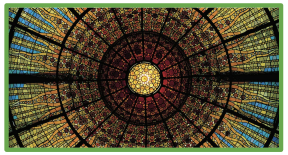
Guiding principles

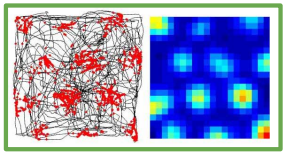
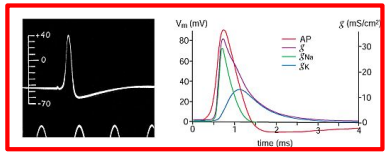
Want to learn more?

The Bitter Lesson, Sutton, incompleteideas.net 2019

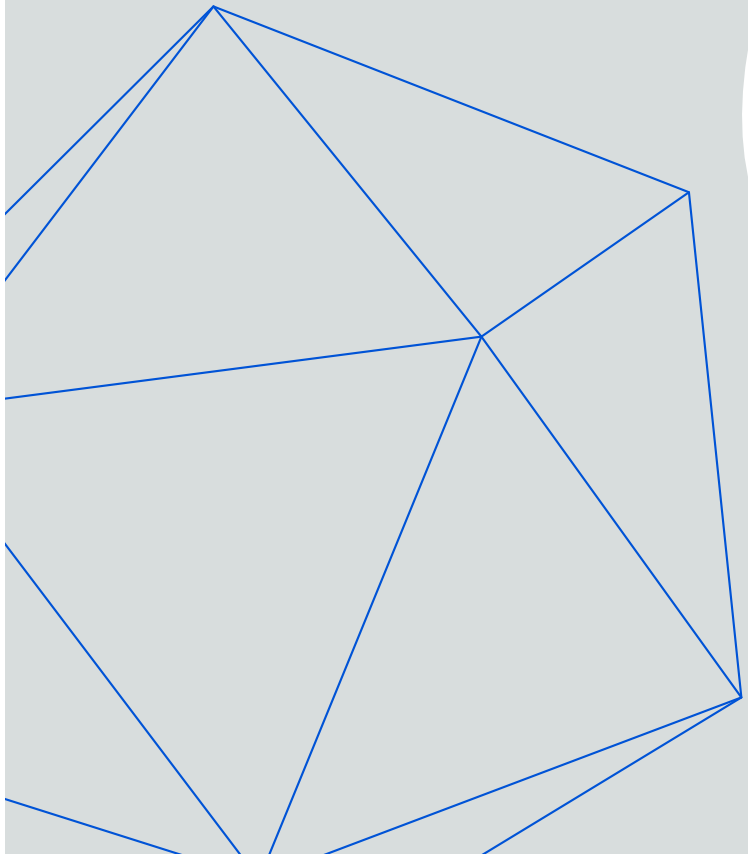
physics → AI ← neuroscience

Avoid details, think about the fundamentals



DeepMind



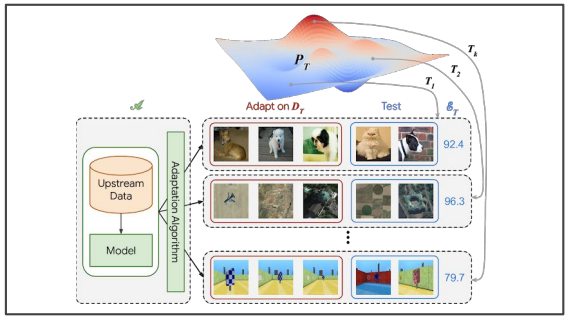
Evaluating representations

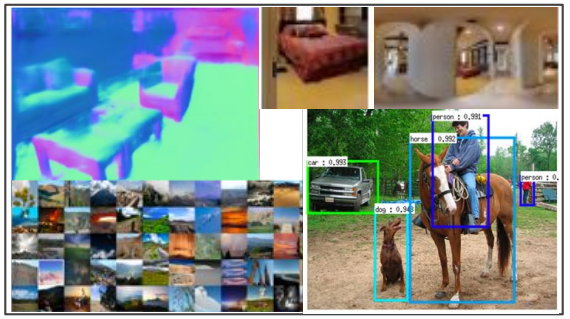
Evaluating representations

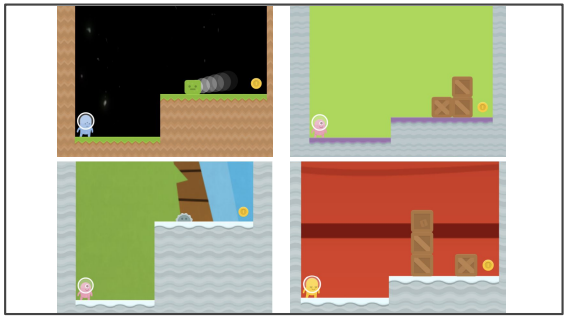
Want to learn more?

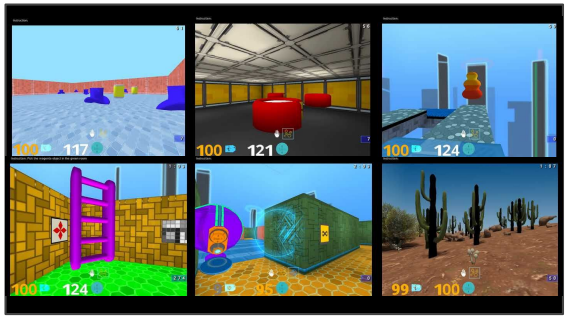
A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark, Zhai et al, arxiv 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties







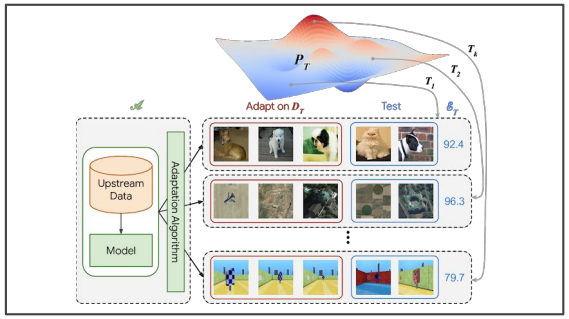


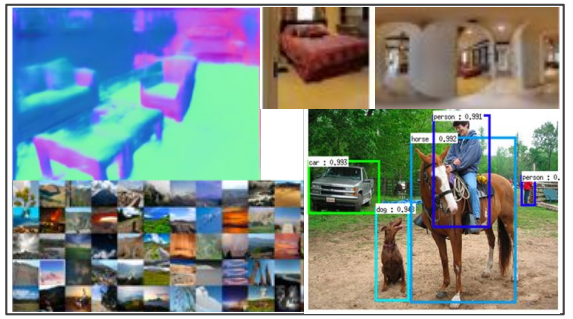
Evaluating representations

Want to learn more?

Scaling and Benchmarking Self-Supervised Visual Representation Learning, Goyal et al, ICCV 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?





Visual Task Adaptation Benchmark (Google)

19 visual tasks split into three groups: natural, specialised and structured. Allowance of 1000 adaptation examples per task.

FAIR Self-Supervision Benchmark (Facebook)

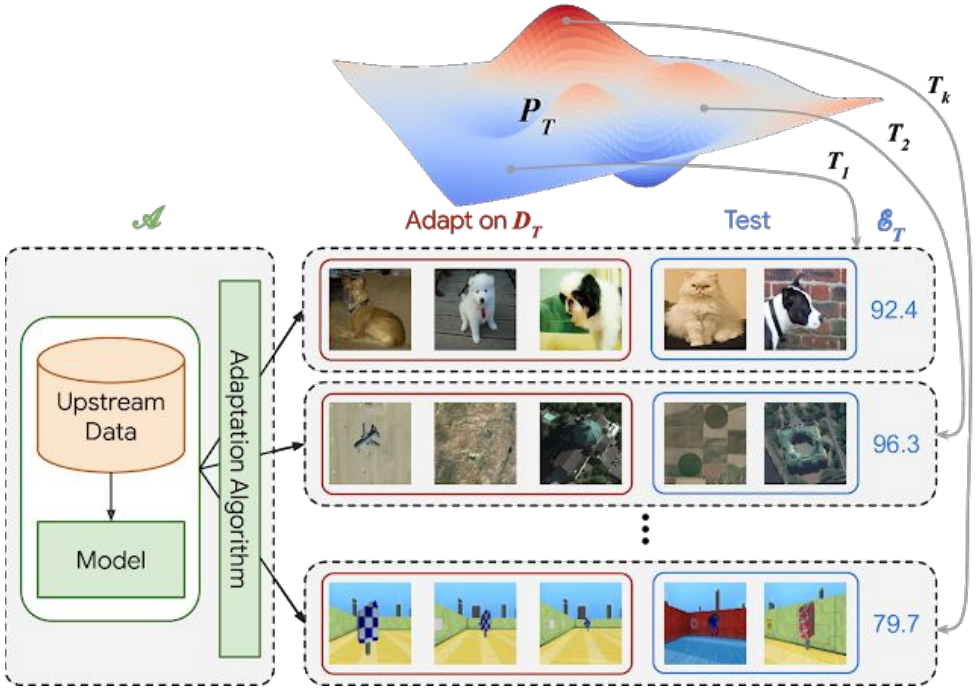
Image classification, object detection, surface normal estimation and visual navigation tasks. Allows limited supervision and fine-tuning.

Evaluating representations

Want to learn more?

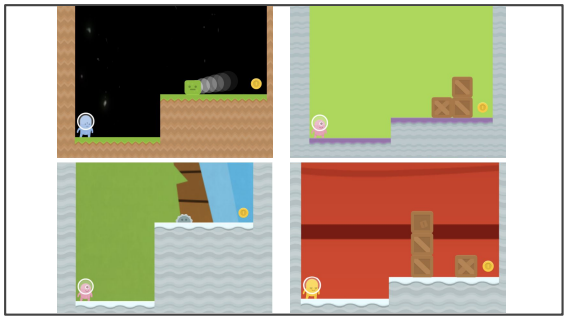
A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark, Zhai et al, arxiv 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?



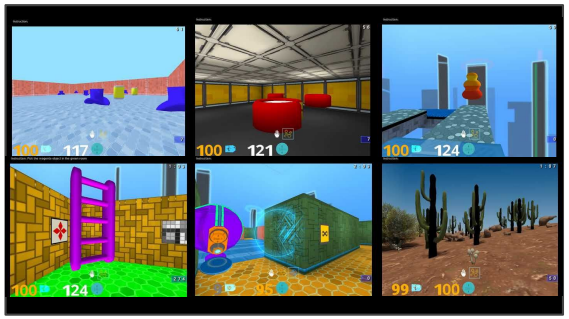
Evaluating representations

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?



CoinRun (OpenAI)

Procedurally generated levels with different degrees of difficulty and a high variability in the game visuals.



DMLab-30 (DeepMind)

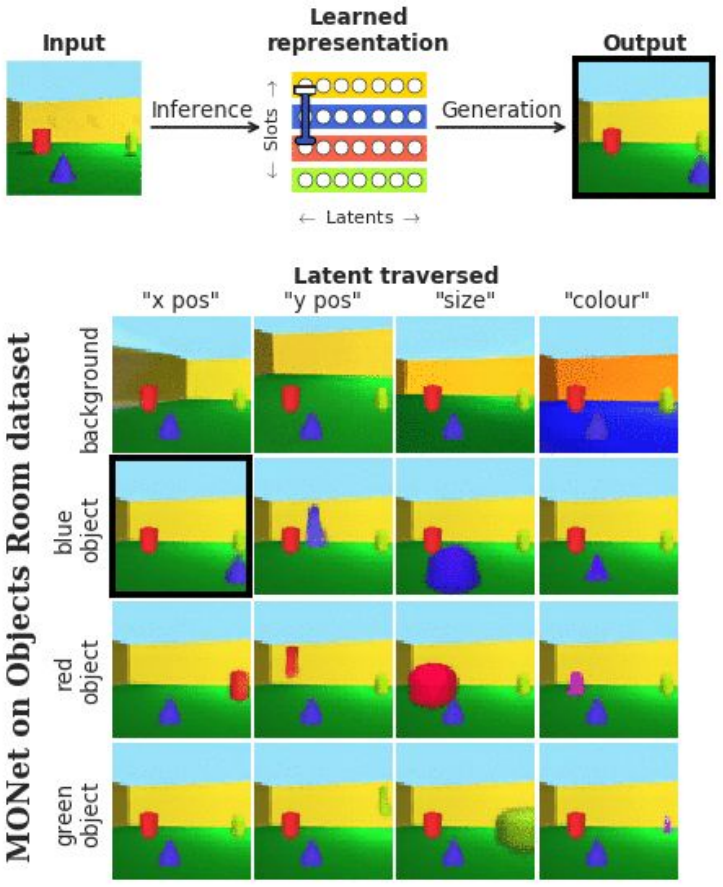
30 varied tasks in a 3D environment, testing navigation, language abilities, multi-agent interactions, long-term planning and more.

Evaluating representations

Want to learn more?

MONet: Unsupervised Scene Decomposition and Representation
Burgess et al., arxiv 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties
  - Latent visualisations

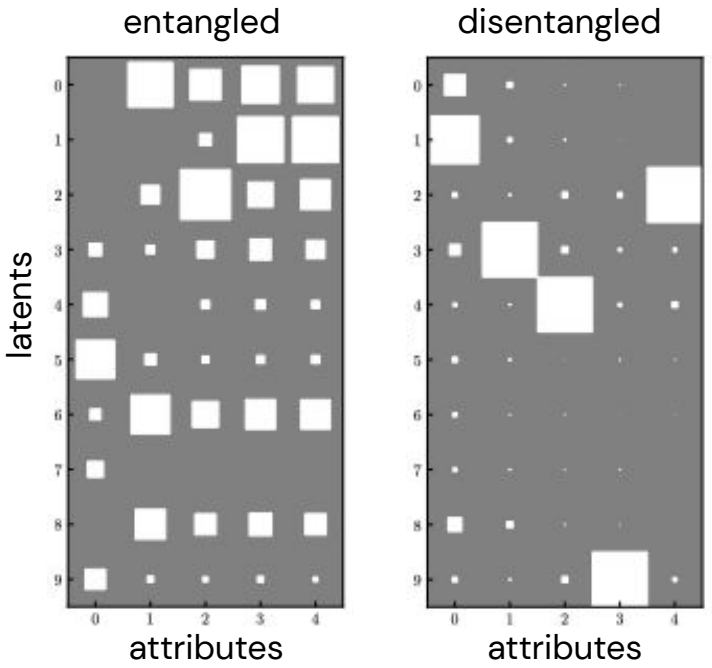


Evaluating representations

Want to learn more?

disentanglement_lib, Bachem & Locatello, github 2019
https://github.com/google-research/disentanglement_lib

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties
  - Latent visualisations
  - Metrics



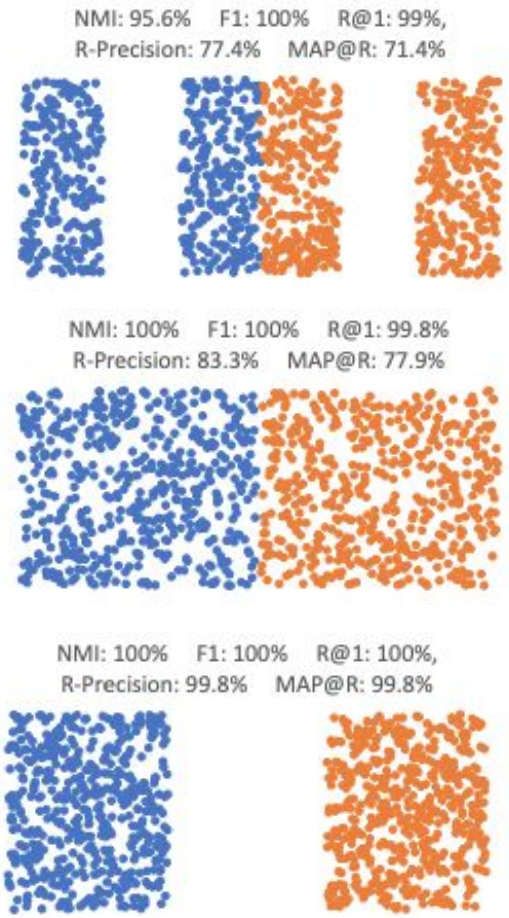
Adapted from Eastwood & Williams (2017)

Evaluating representations

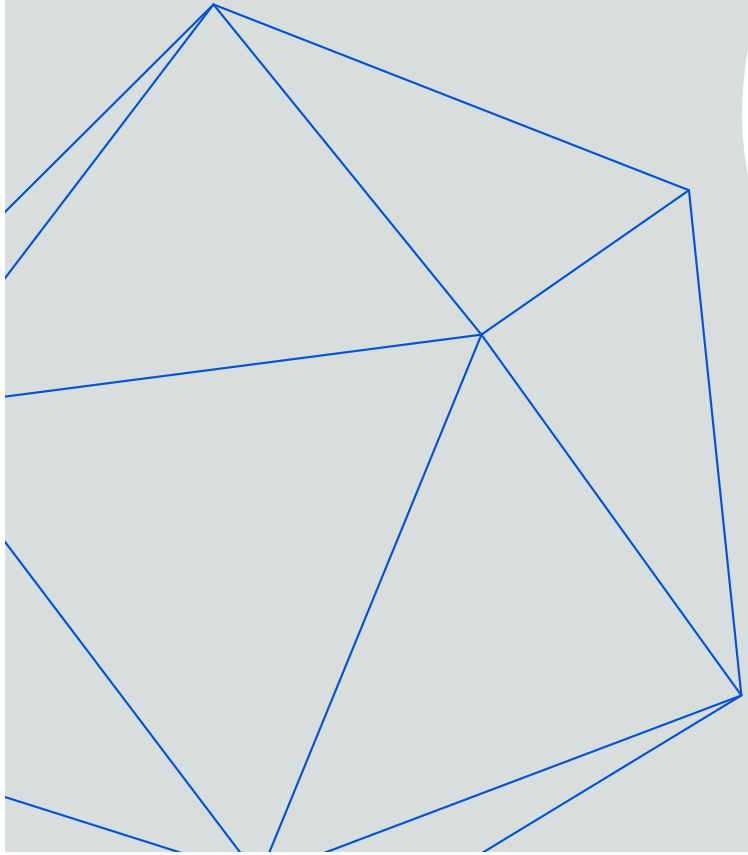
Want to learn more?

A Metric Learning Reality Check, Musgrave et al, arxiv 2020

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties
  - Latent visualisations
  - Metrics



DeepMind



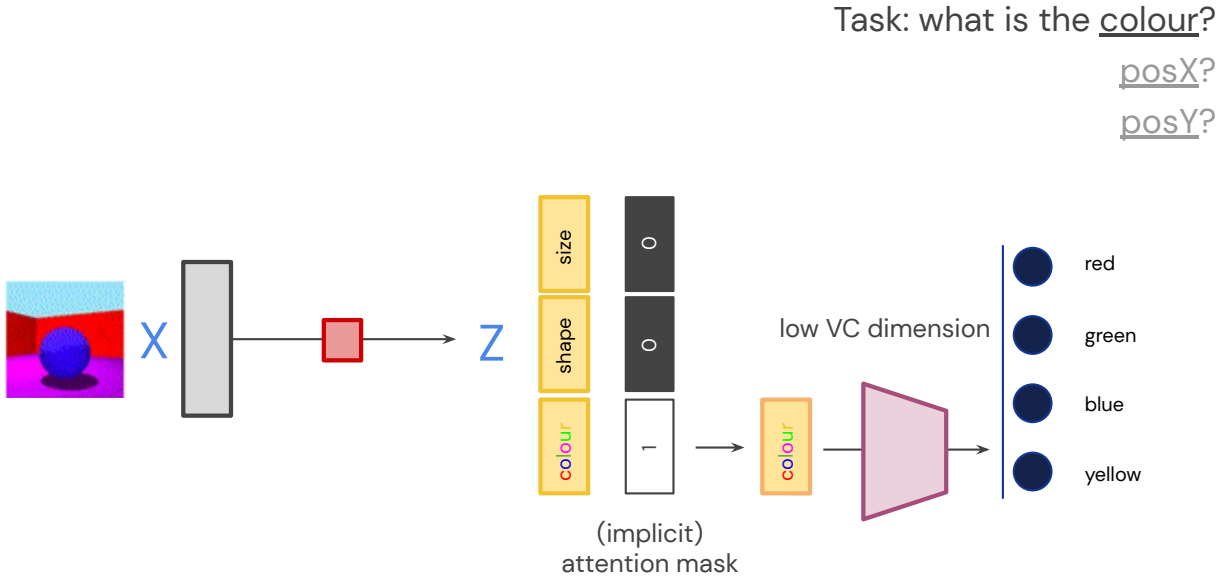
Utility of learning better representations

Utility of disentangled representations

Want to learn more?

Deep Learning of Representations: Looking Forward, Bengio, SLSP 2013

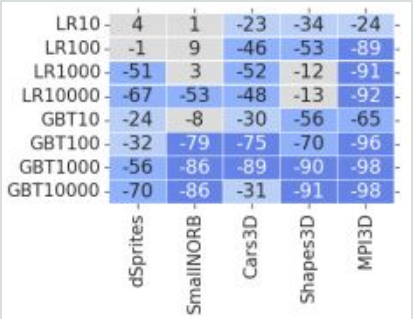
- Data efficiency
- Generalisation under covariate shift
- Fairness
- Abstract reasoning



Utility of disentangled representations

Want to learn more?

Weakly-Supervised Disentanglement Without Compromises, Locatello et al, arxiv 2020

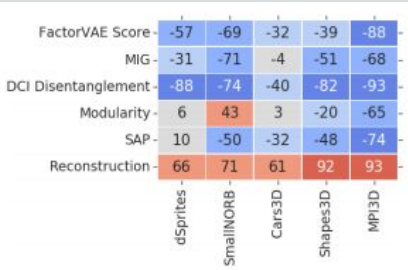


Data efficiency

Predict values of generative factors from representation in **low data regime**

Using logistic regression (LR) or gradient boosted trees (GBT)

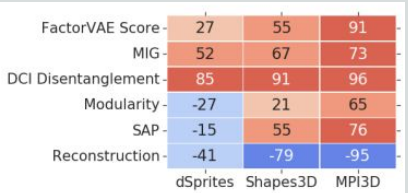
Higher disentanglement correlates with better **accuracy**



Fairness

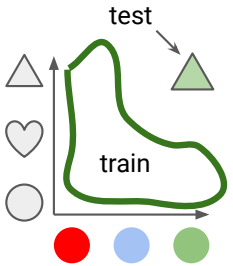
Higher disentanglement correlates with classifiers that are **fairer to unobserved sensitive variables** independent of the target variable

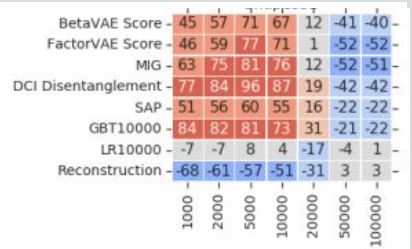
Use GBT10000 classifier



Generalisation

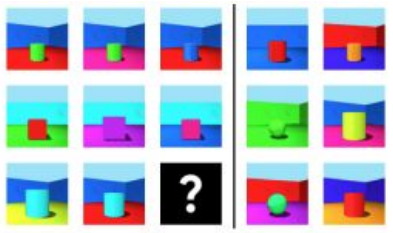
Higher disentanglement correlates with **generalisation under covariate shifts**





Abstract reasoning

Higher disentanglement correlates with **accuracy on abstract visual reasoning tasks in lower data regimes**

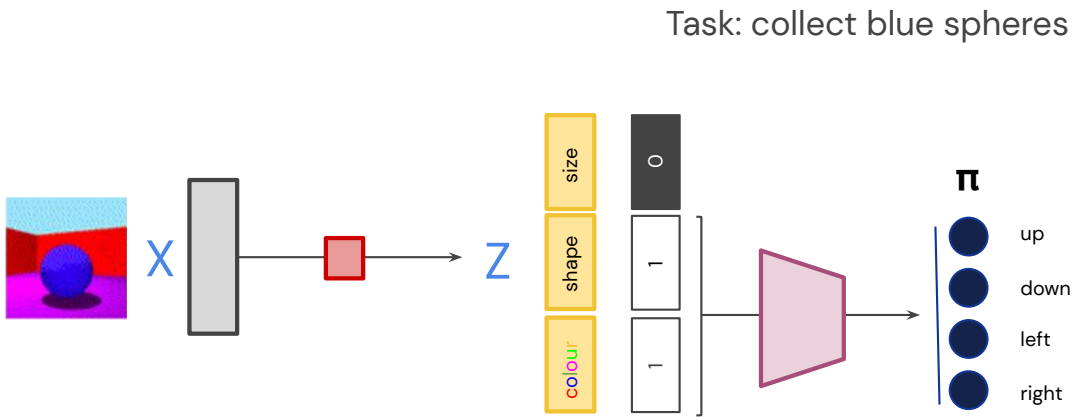


Utility of disentangled representations

Want to learn more?

DARLA: Improving Zero-Shot Transfer in Reinforcement Learning, Higgins, Pal et al, ICML 2017

- Data efficiency
- Generalisation under covariate shift
- Fairness
- Abstract reasoning
- Transfer

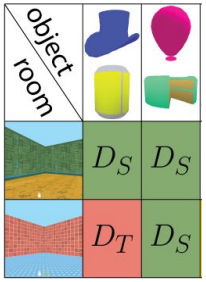


Utility of disentangled representations

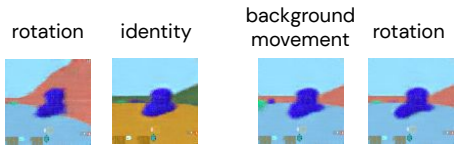
Want to learn more?

DARLA: Improving Zero-Shot Transfer in Reinforcement Learning, Higgins, Pal et al, ICML 2017

- Data efficiency
- Generalisation under covariate shift
- Fairness
- Abstract reasoning
- Transfer



Vision Type	DQN	DeepMind Lab A3C	EC
Baseline agent	1.86 ± 3.91	5.32 ± 3.36	-0.41 ± 4.21
UNREAL	-	4.13 ± 3.95	-
DARLA$_{FT}$	$\mathbf{13.36 \pm 5.8}$	1.4 ± 2.16	-
DARLA$_{ENT}$	3.45 ± 4.47	15.66 ± 5.19	5.69 ± 3.73
DARLA$_{DAE}$	7.83 ± 4.47	6.74 ± 2.81	5.59 ± 3.37
DARLA	10.25 ± 5.46	$\mathbf{19.7 \pm 5.43}$	$\mathbf{11.41 \pm 3.52}$



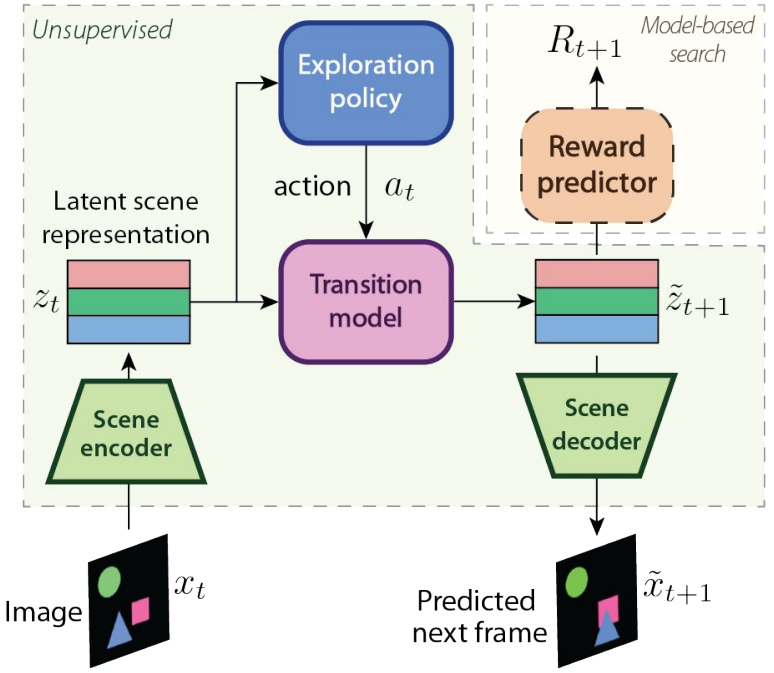
Utility of object-based representations

Want to learn more?

COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration, Watters, Matthey et al, arxiv 2019

During unsupervised exploration stage learn:

- Object decomposition and feature disentangling
- Action-conditioned transition model of the environment
- Curiosity-driven exploration policy



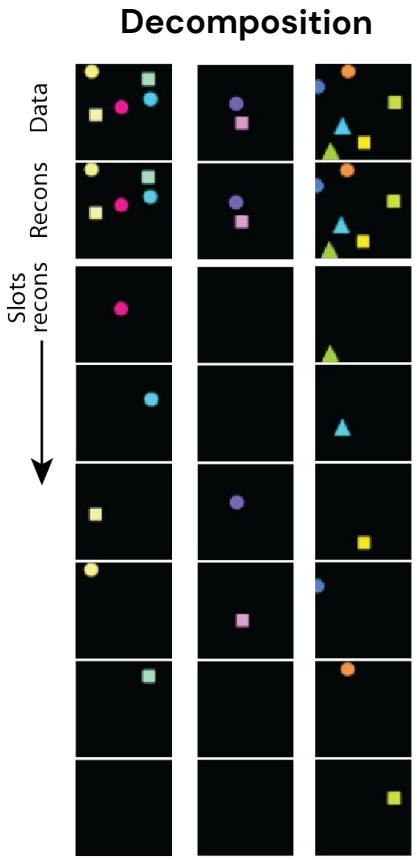
Utility of object-based representations

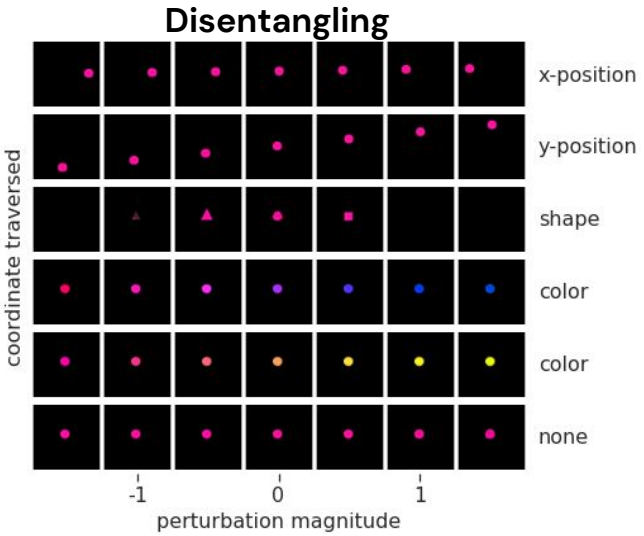
Want to learn more?

COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration, Watters, Matthey et al, arxiv 2019

During unsupervised exploration stage learn:

- Object decomposition and feature disentangling
- Action-conditioned transition model of the environment
- Curiosity-driven exploration policy





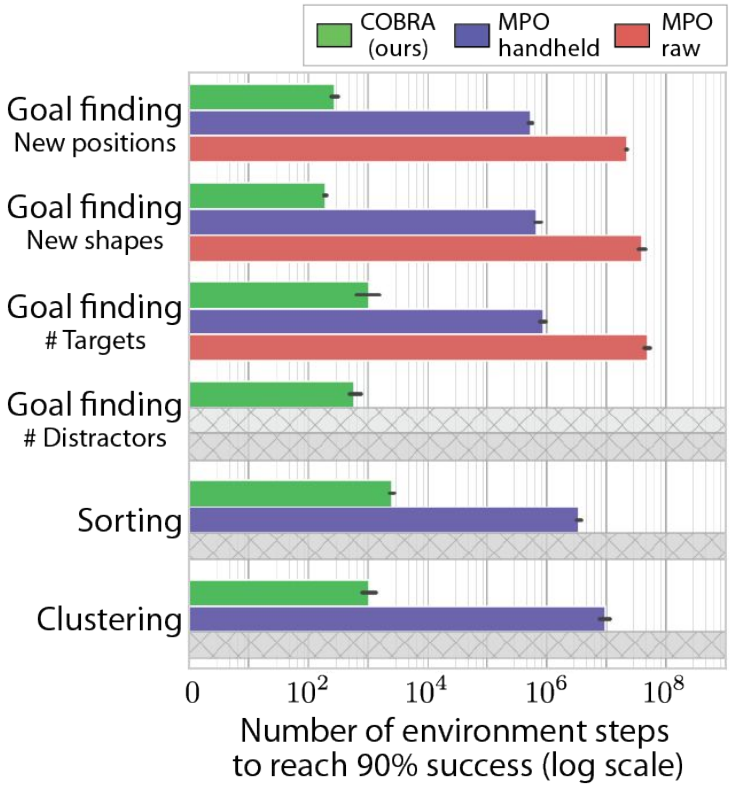
Utility of object-based representations

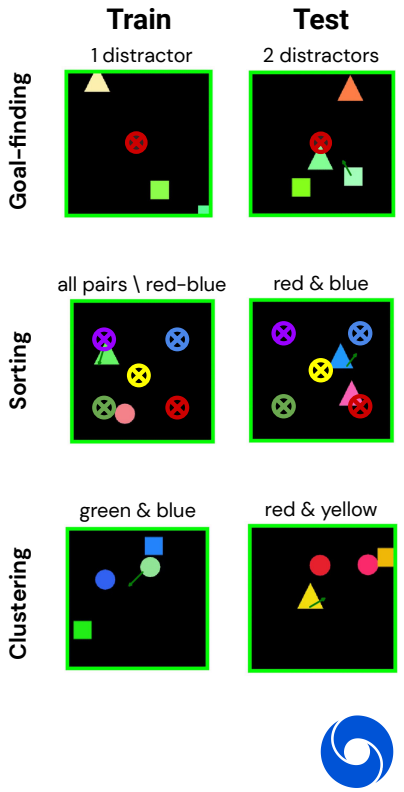
Want to learn more?

COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration, Watters, Matthey et al, arxiv 2019

Task learning reduced to to learning reward function for model-based search.

- Better data efficiency
- Better generalisation



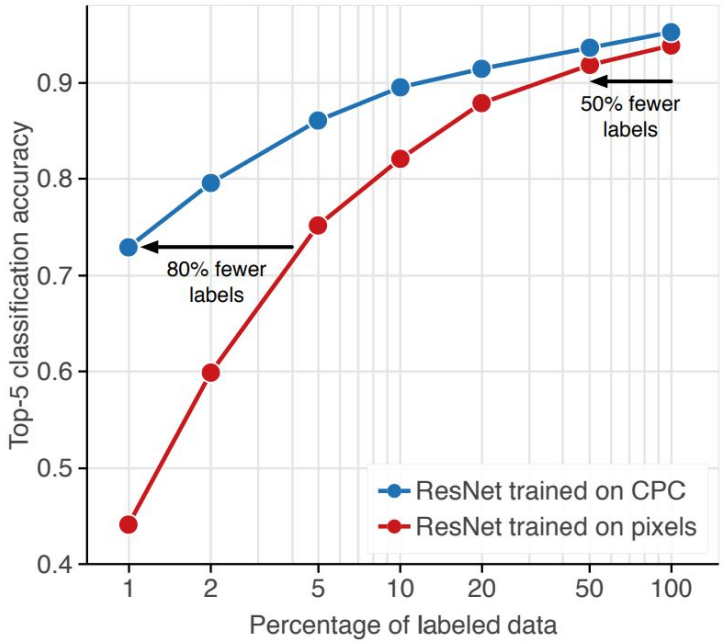


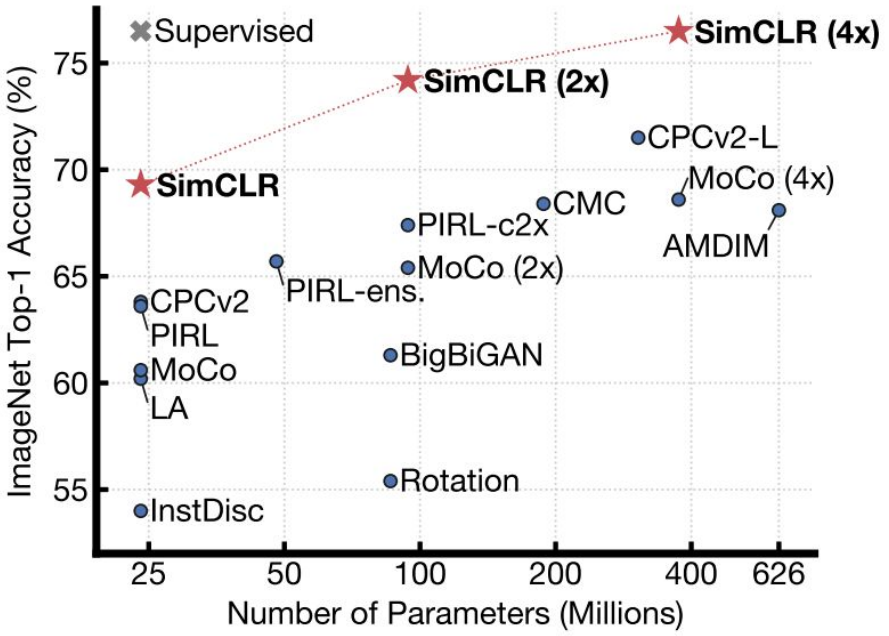
Utility of contrastive methods

Want to learn more?

Data-Efficient Image Recognition with Contrastive Predictive Coding, Olivier J. Hénaff et al, ICML 2020

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML 2020



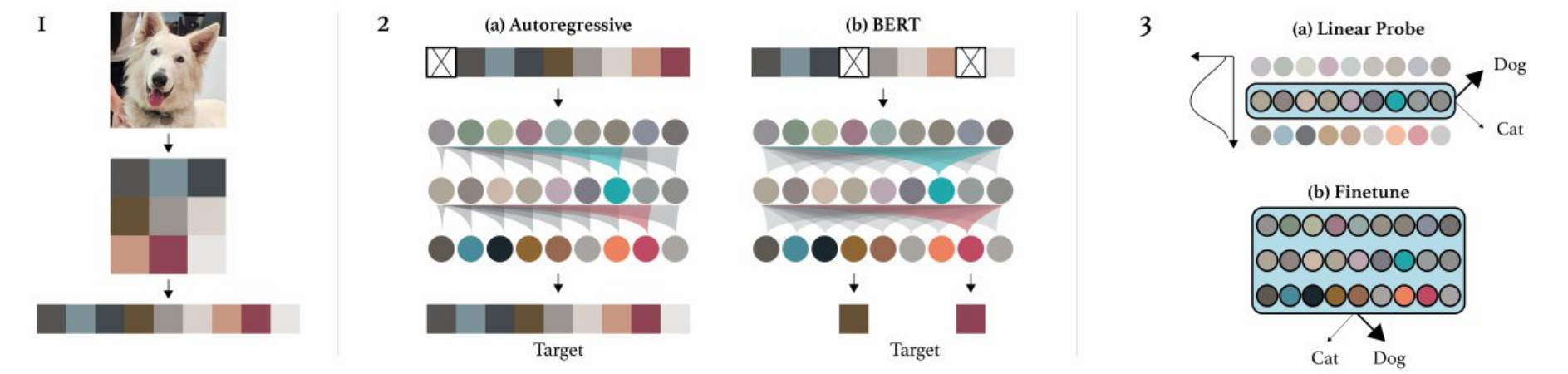


Utility of attention-based methods

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML 2020

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al, NAACL 2019

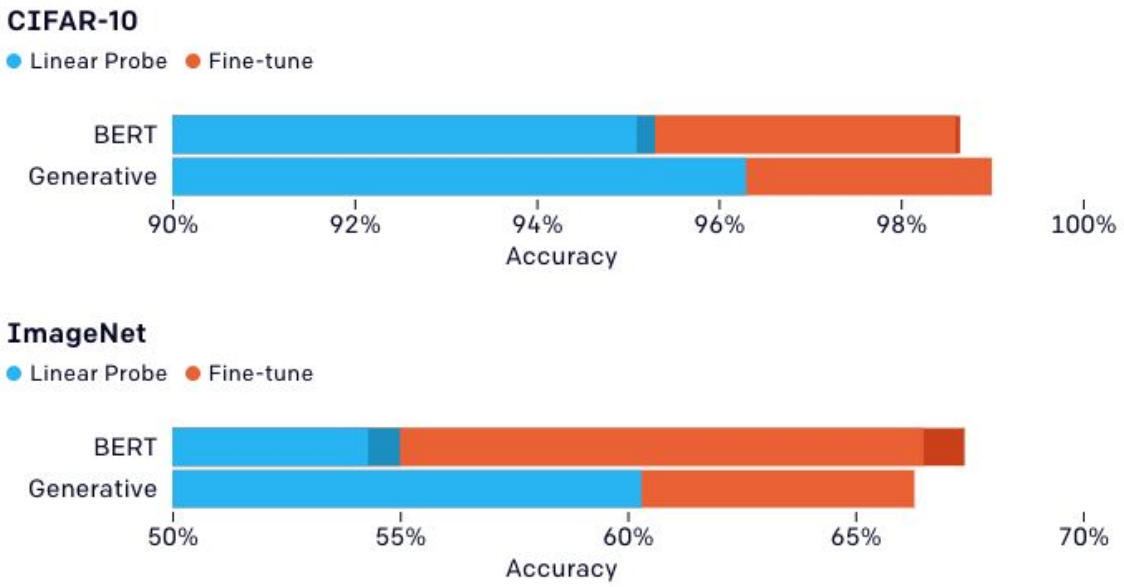


Utility of attention-based methods

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML 2020

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al, NAACL 2019



Utility of attention-based methods

Want to learn more?

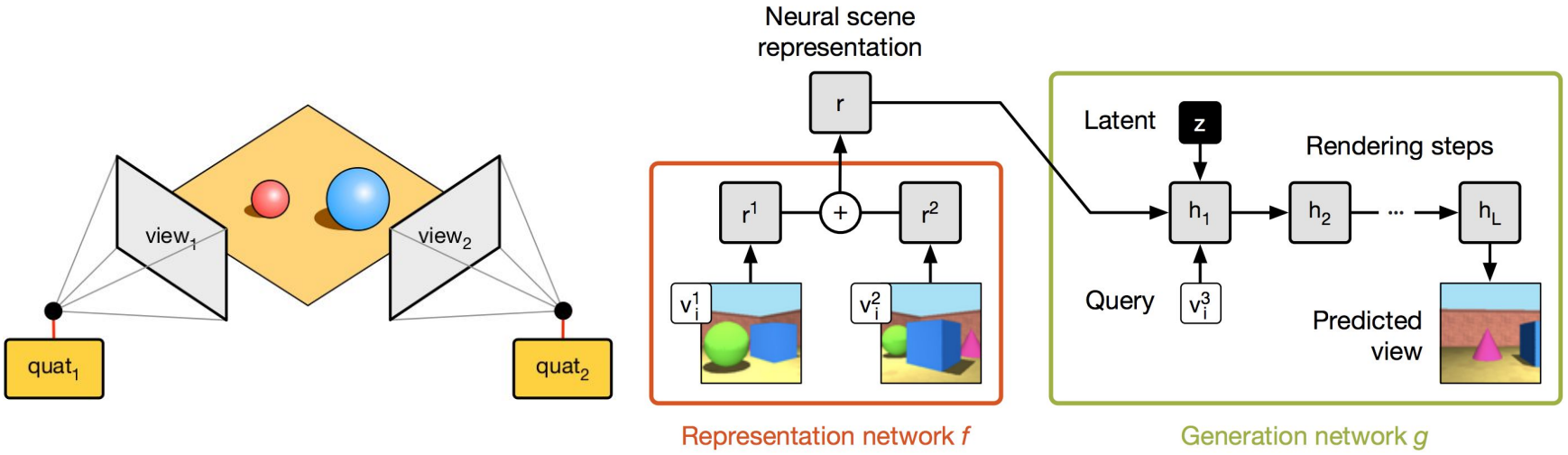
Generative Pretraining from Pixels, Chen et al, ICML 2020

EVALUATION	MODEL	ACCURACY	PRE-TRAINED ON IMAGENET W/O LABELS	PRE-TRAINED ON IMAGENET W/ LABELS
CIFAR-10 Linear Probe	ResNet-152[50]	94.0		✓
	SimCLR[12]	95.3	✓	
	iGPT-L 32x32	**96.3**	✓	
CIFAR-100 Linear Probe	ResNet-152	78.0		✓
	SimCLR	80.2	✓	
	iGPT-L 32x32	**82.8**	✓	
STL-10 Linear Probe	AMDIM-L[13]	94.2	✓	
	iGPT-L 32x32	**95.5**	✓	
CIFAR-10 Fine-tune	AutoAugment[51]	98.5		
	SimCLR	98.6	✓	
	GPipe[15]	**99.0**		✓
	iGPT-L	**99.0**	✓	
CIFAR-100 Fine-tune	iGPT-L	88.5	✓	
	SimCLR	89.0	✓	
	AutoAugment	89.3		
	EfficientNet[52]	**91.7**		✓

Utility of better belief representations (GQN)

Want to learn more?

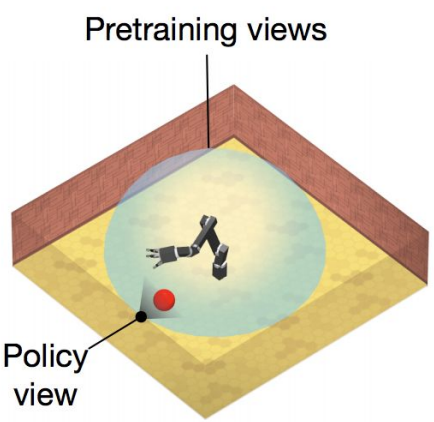
Neural Scene Representation and Rendering, Eslami et al, Science 2018

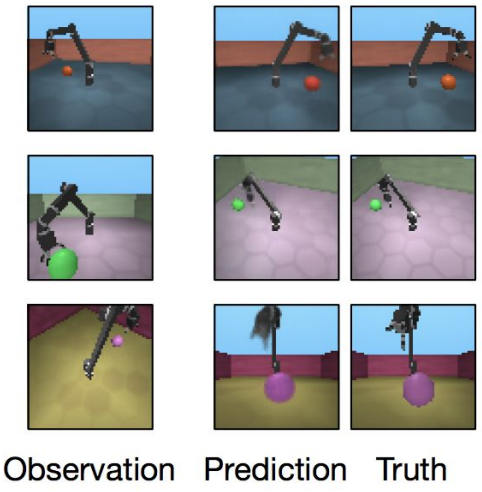


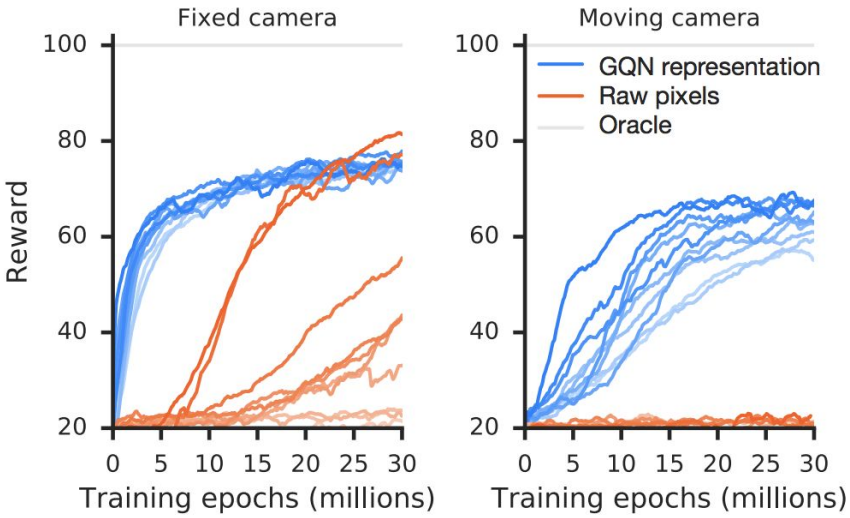
Utility of better belief representations (GQN)

Want to learn more?

Neural Scene Representation and Rendering, Eslami et al, Science 2018



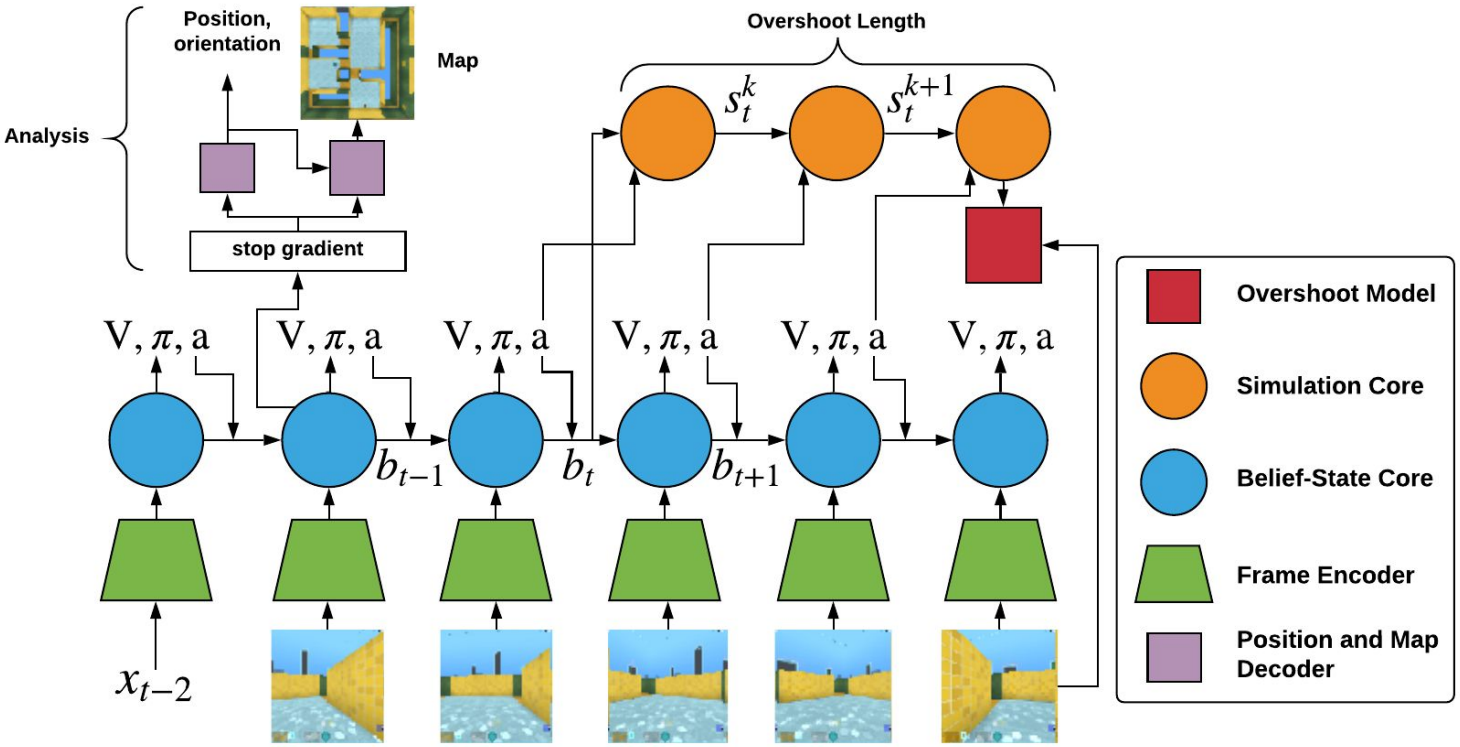




Utility of better belief representations (SimCore)

Want to learn more?

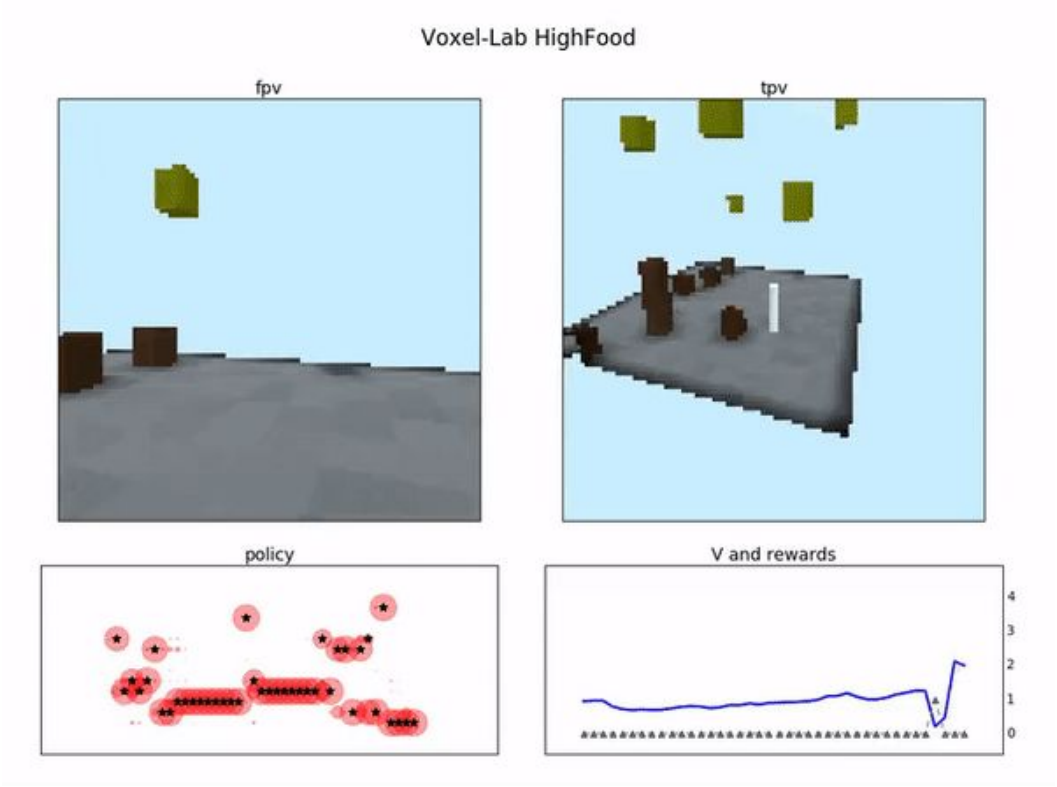
Shaping Belief States with Generative Environment Models for RL, Gregor et al, NeurIPS 2019

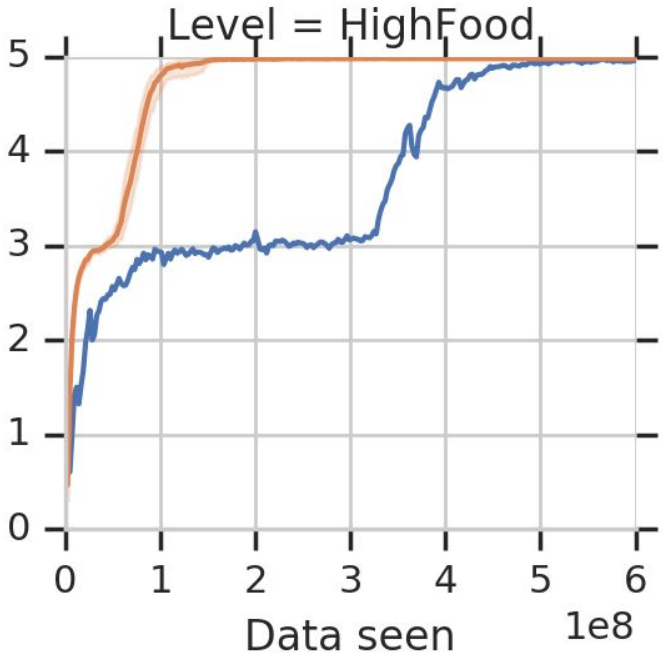


Utility of better belief representations (SimCORE)

Want to learn more?

Shaping Belief States with Generative Environment Models for RL, Gregor et al, NeurIPS 2019

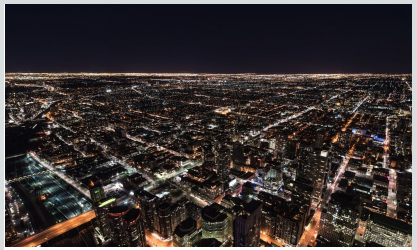




Take away messages

01

Representations are **useful abstractions** that make downstream computations and tasks more **efficient**.

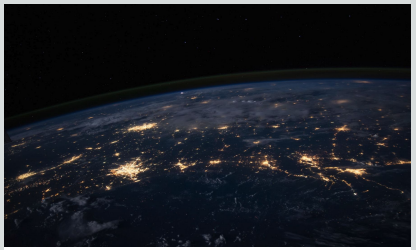


02

Representation learning problem is **under-specified**.

Current approaches tend to tradeoff **generality vs interpretability.**

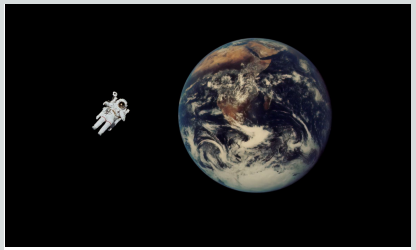
Recent progress is surprisingly impressive.



03

The **diverse "model zoo"** can be understood using **simple theoretical taxonomy**.

Thinking about **density modelling**, **manifolds** and the level of **modelling detail** may help understand the tradeoff and areas of improvement for different methods.



04

Multi-disciplinary insights may help resolve some of the under-specification.

Different representations **help with different tasks**.

No agreed upon **evaluation method**.

video credit: Francis Vachon
www.francisvachon.com

Thank you


Questions